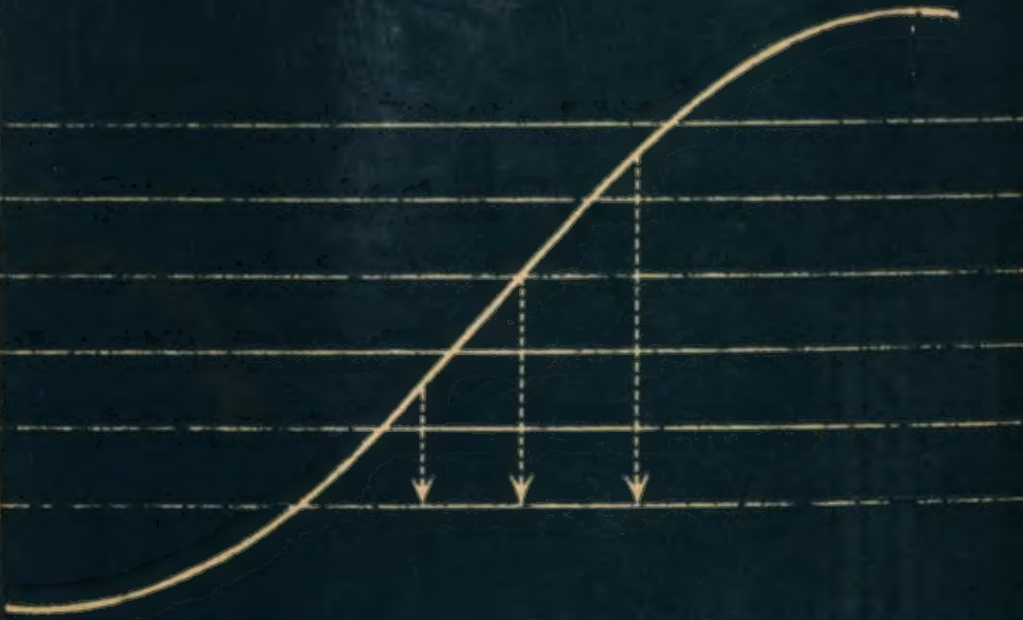
Business Statistics



N. K. NAG
J. C. MAITY

BUSINESS STATISTICS

[*For B. Com. (Hons.) Class*]

N. K. NAG

HEAD OF THE DEPARTMENT OF MATHEMATICS, UMESCHANDRA COLLEGE ; LECTURER, DEPARTMENT OF MATHEMATICS AND STATISTICS, CITY COLLEGE OF COMMERCE AND BUSINESS ADMINISTRATION, CALCUTTA

AND

J. C. MAITY

HEAD OF THE DEPARTMENT OF BUSINESS MATHEMATICS AND STATISTICS, CITY COLLEGE OF COMMERCE AND BUSINESS ADMINISTRATION, CALCUTTA

nbs

THE NEW BOOK STALL
BOOK-SELLERS, PUBLISHERS & EXPORTERS
5/1, Ramanath Majumdar Street, Calcutta-700 009

First Edition : November, 1981

Price : Rs. 32·50

PUBLISHED BY SRI MAHENDRA NATH PAUL,
THE NEW BOOK STALL, 5/1, RAMANATH MAJUMDAR STREET,
CALCUTTA-700 009 AND PRINTED BY SRI TRIDIBESH BASU,
AT K. P. BASU PRINTING WORKS, 11, MOHENDRA GOSSAIN LANE,
CALCUTTA-700 006

PREFACE

This book (in two volumes) is primarily meant for Two Year Degree Course for B. Com. (Hons.) students under new syllabus of Calcutta University. There are many students in B. Com. (Hons.) Course who have passed previous *Plus Two* Course without taking Mathematics as one of the subjects. This book is prepared keeping those students in mind.

Different chapters in the prescribed syllabus have been expounded with great care with the help of explanatory notes followed by suitable examples. For the guidance of all types of students quite a large number of hard examples have been worked out in all the chapters. Advanced exercises have been provided for the benefit of ambitious students. While writing the book, we have always kept in mind the nature and types of problems usually set in different examinations.

Apart from University examinations, few essential questions set in C.A. and I.C.W.A. of India Examinations have been given in the book for the interest of students. The authors thank those institutions mentioned for the kind permission given to make use of such questions.

In preparing this book we had to consult many books written by Indian and foreign authors. Professors D. Roy of Gobardanga Hindu College, G. S. Mukherjee and K. Debnath of Umeschandra College, N. L. Lahiri and R. K. Bhattacharjee of City College of Commerce encouraged and helped us in many respects. We express our indebtedness and heartful gratitude to all of them.

Sri Subhash Chandra Ghosh, an ex-student, took an active part in laying out the Charts, Tables and also made through proof-reading while printing the book. We are pleased to record his untiring effort.

The authors are also thankful to Sri Mahendra Nath Paul of The New Book Stall and the authorities of K. P. Basu Printing Works.

In spite of best efforts, some mistakes might have been crept in. We shall be highly obliged if any reader kindly brings such mistakes to our notice. Suggestions, if any, for the improvement of the book will be gratefully acknowledged.

CALCUTTA
Nov., 1981

AUTHORS

CONTENTS

BUSINESS STATISTICS

1

INTRODUCTION

The word STATISTICS was originally, in the earlier age, used for collection and arrangements of facts, not necessarilly numerical type, about a state or the people belonging to a state. Now a days the word STATISTICS is generally used for making census operations, collecting information regarding social and economic status of different people, of the different part of a country.

The theory of probability—the basic principle of statistics was first discussed by G. Cardano. The theory is recognised to-day as one of the fundamental laws of statistics and statistical conclusions are largely based on it. Of course, the studies were extended by Gauss, Bayes Euler, Lagrange, Hain, Knapp and Lexis—only few names to mention.

Statistics in India.

In India Statistics came during the reigns of Ashoka, Gupta Dynasty and Mughal rules. We find Kautilya Artha-shastra is replete with statistics of land, prices, wages, of population etc., collected during Maurya rule. It may be noted Todar Mal, the Finance Minister of Akbar compiled statistics of land, agriculture, trade etc. for placing the land revenue and tax system in a systemetic order.

In the second half of 18th century East India Company also started collecting statistics mainly on agriculture. In the publication of 'Statistical Abstract of British India' during British Government reign, we find a lot of statistical informations.

Sense.

The word statistics is now used in two different senses :—

(i) Statistics as a *plural* noun, mean a collection of numerical facts (statistics of birth or death) or derivation of numerical facts *i.e.* percentages, averages, estimation regarding any population etc.

(ii) Statistics, as a *singular* noun, on the other hand, refers, to various methods called statistical methods, adopted for collection, analysis and interpretation of numerical data.

It should be noted, here, that a single and unconnected figure cannot be called as statistics. A single figure is incapable of comparison analysis or interpretation. There should be at least two figures. Further the figures, in question, should be capable of being placed in relation to each other.

In whatever sense the word statistics is used, it should be always remembered that the subject is mainly concerned with facts expressed in a numerical form *i.e.* with quantitative details and not with the qualitative descriptions.

Relations.

Statistics and Mathematics. During the 17th century, the methods of statistical science were used under the name of *Political Arithmetic.* In 18th century, a relation between statistics and mathematics was formed on the basis of the theory of probability when Jacob Bernoulli (1654—1705) stated the 'Law of large numbers' in his great work *Ars Conjectandi* published eight years after his death. L. A. J. Quetlet (1796—1874) also emphasised the importance of 'Law of large numbers'. Daniel Bernoulli (1700—1782) laid a solid foundations on the theory of probability. On these foundations laid by the mathematicians (mentioned a few only), modern theory of statistics was gradually built up.

Statistics and Economics. The relationship between these two sciences became intimate rather late, although a reference of relationship was made by Sir William Petty in his work, Political Arithmetic published in 1690. By the 18th century, statistical data relating to population, taxes, agriculture, industry, trade etc. used to be collected in most civilised countries, but there was no relationship between statistical information and economic theory. In 1871, W. S. Jevons wrote in his *Theory of Political Economy* that 'the deductive science of economy must be verified and rendered useful from the purely inductive science of statistics. Theory must be invested with the reality and life of fact. Political economy could gradually be developed into an exact science, if commercial statistics were far more complete and precise." He developed the technique of analysis of time series. Rightly he has been called the 'Father of Index Numbers'. Besides Jevons, the Historical School (1843—1883) brought Statistics and Economics more closer. In fact Roscher, Knies, Hilderbrand and Cliff Leslie believed that economic doctrines should not be argued in the abstract, but to be inductively proved. This effect was indeed great and the science of economics no more remained deductive in approach. By the end of 19th century, attitude of economists towards the inductive method had become friendly. In 1907, Alfred Marshell wrote that disputes as to the methods of study in economics had ceased, that qualitative analysis had performed the greater part of its work and the progress in the quantitative analysis depended upon the growth of realistic statistics. He explained that induction and deduction were both needed for scientific thought, as the right and left feet were both needed for walking.

Since 1890, two factors have brought about the fundamental change in the relationship of statistics and economics. The first is the development of statistical methods, like probability sampling, correlations, periodicity and index number etc., secondly there is an enlargement of statistical data in recent years. During the period, eminent statisticians like August Meitzen, Karl Pearson, G. U. Yule, C. V. Davenport, A. L. Bowley, W. Persons, R. A. Fisher etc. have made valuable contributions regarding the development of the science,

The improvement of statistical methods and enlargement of statistical data have thus brought statistics and economics very close to each other.

Limitations of the Science of Statistics.

1. *Statistics studies only quantitative phenomena.* One of the important limitations of the science is that it deals only with those phenomena which can be expressed by quantity. And phenomena which cannot be expressed by figure, like brave, honesty, intelligency etc. are of little use. Of course efficiency, intelligence etc. can be compared on the basis of marks obtained, but still these are only indirect method of approach.

2. *Statistical laws are true only on an average.* It is known that the laws of physics, mathematics, chemistry etc. are exact and universally applicable. The laws of statistics are not so. Statistics deals with such phenomena which are affected by multiplicity of causes and it is not possible to study the effects of each of these factors individually as is done by experimental methods. Due to this limitation, the results obtained are not perfectly accurate rather approximation.

3. *Statistics deals with aggregates and not with individuals.* Statistics deals with aggregate, although these aggregates are often reduced to single figures for analysis. A series of figures is condensed into an average for comparison, but an individual item of the same series has no specific recognition.

4. *It is liable to be misused.* Any person can misuse statistics and draw any type of conclusions he desires. Statistical methods can be properly used by only those who have a sound knowledge about them and their use by less expert hands is sure to give inaccurate results.

Those who use statistics must be aware of the limitations. According to W. I. King, "*statistics is a most useful servant but only of great value to those who understand its proper use.*"

Characteristics.

(i) Statistics must be quantitatively expressed : Phenomena which cannot be expressed numerically like intelligence, honesty, brave etc. ; are of little use in statistics. Qualitative expression like young, middle-aged or old should be expressed by say 25, 45 or 65 years.

(ii) Statistics are always aggregates : Single and unconnected figures are not statistics as they cannot be studied in relation to each other. A single age of 20 or 25 years is not statistics but series of ages. A single birth or sale is not a statistics while a number of births or sales are so, since they can be studied in relation to time or place.

(iii) Relation to enquiry : The significance of certain figures can be better appreciated when they are compared with others of the same type.

(iv) Relation to each others : Statistics are generally collected to facilitate comparison in point of time or place or condition. If, however, the collected data are unfit for comparison, then much of the importance is over. For this purpose, the figures should be of same nature. For instance, ages of husbands are to be compared only with the corresponding ages of wives and not with the lengths of the trees.

Division of Works.

The whole work of a statistician can be broadly divided as follows :

(i) Collection of data

(ii) Classification and tabulation

(iii) Analysis

(iv) Interpretation

2

COLLECTION OF DATA

A statistician begins the work with the collection of data *i.e.* numerical facts. The data so collected are called *raw materials* (*or raw data*). It is from these raw materials, a statistician analyses after proper classification and tabulation, for the final decision or conclusion. Therefore it is undoubtedly important that the raw data collected should be clear, accurate and reliable.

Before the collection of data, every enquiry must have a definite object and certain scope, that is to say, what information will be collected, for whom it will be collected, how often or at what periodicity it will be collected and so on. If the object and the scope of enquiry are not clearly determined before hand, difficulties may arise at the time of collection, which will be simply a wastage of time and money.

Statistical Unit.

The unit of measurement applied to the data in any particular problem is the statistical unit.

Physical units of measurement like quintal, kilogramme, metre, hour and year etc. do not need any explanation or definition. But in some cases statistician has to give some proper definition regarding the unit. For *example*, the wholesale price of commodity. Now what does the form 'wholesale price' signify ? Does it stand for the price at which the producer sells the goods concerned to the stockist, or the price at which the stockist sells to a wholesaler ? Is it the price at which the market opened at the day of enquiry ? Many such problems may arise as stated. It is thus essential that a statistician should define the units of data before he starts the work of collection.

Requirements.

1. Its definition must be unambiguous, simple and complete itself. If the unit is not definite, the data collected might be inaccuracy. So it is necessary that the units should be properly defined.

2. It should be stable in character. If there is any fluctuation, the data cannot be compared. For example, if one seer is equal to

0'92 Kg. at one place, at one place '90 Kg. and at other 1 Kg. the data collected can never be compared.

3. The unit should be homogeneous *i.e.*, the unit should imply the same characteristic and also should be uniform throughout the enquiry. If the data are not homogeneous, comparison cannot be made. Now if the data are heterogeneous, they may be broken up into small homogeneous classes. For example, if the data relating to failures in a certain examination are being collected, then the failures can be divided into a number of classes on the basis of subject wise or absentees (total or partial).

4. The statistical unit should be appropriate to enquiry.

Types.

The units are of two types :—

1. Units of collection.
2. Units of analysis and interpretation.

Units of collection are those units in terms of which measurements are made (or estimated). Production of cotton textile industry in bales, consumption of electricity in kilowatts, production of a cereal in a ton etc. are the examples in simple unit. And ton-mile, industrial accident, credit sale etc. are the examples in composite unit. Here tone-mile indicate the number of tons multiplied by the number of miles carried. It is a combination of qualifying word to a simple unit.

Units of analysis are those units in terms of which data are compared. They include ratios, percentages, rates etc. All these are useful for comparison.

Degree of Accuracy.

Another point that should be decided in advance is the degree of accuracy to which the data is to be collected. For instance, it should be decided whether prices or wages are to be quoted correct to paise or rupee. Similarly the weight to gramme or kilogramme. The aims should be, to determine the unit in advance clearly and precisely, so as to avoid the wastage of time and labour.

Types and Methods of Collection of Data.

Statistical data are usually of two types, (i) *Primary* (ii) *Secondary*.

Data which are collected for the first time, for a specific purpose are known as primary data, while those used in an investigation, which have been originally collected by some one else, are known as secondary data.

For *example*, data gathered by the Office of the Registrar General of Census Operation are primary data, but are secondary when used by others.

On the basis of primary and secondary data the methods of collecting statistical data have been divided into *Primary method* and *Secondary method*.

Distinguish between Primary and Secondary Data.

1. Primary data are those data which are collected for the first time and thus original in character. Secondary data are those data that have already been collected earlier by some other persons.

2. Primary data are in the form of raw materials to which statistical methods are applied for the purpose of analysis. On the other hand, secondary data are in the form of finished products as they have been already statistically applied.

3. Primary data are collected directly from the people to which enquiry is related. Secondary data are collected from published materials.

4. If observed closely the difference is of one degree only. Data are primary to an institutions collecting it, while they are secondary for all others. Thus data which are primary in the hands of one, are secondary in the hands of other.

Primary Method.

The following methods are common in use :—

(i) ***Direct Personal Observation.*** Under this method, the investigator collects the data personally. He has to go to the spot for conducting enquiry and has to meet the persons concerned. It is essential that the investigator should be polite, tactful and has a sense of observation.

This method is applicable when the field of enquiry is small and there is an intention of greater accuracy. This method however, gives satisfactory rasult provided the investigator is fully dependable.

(ii) ***Indirect Oral Investigation.*** In this method data are collected through indirect sources. Persons having some knowledge regarding the enquiry, are cross-examined and the desired information is collected. Evidence of one person should not relied, but a number of views should be taken to find out real position. This method is usually adopted by enquiry committees or commissions appointed by governments or semi-governments or private institutions. Certain precautions are to be taken here. Firstly it should be seen whether the informant knows full facts of the problem under investigations. Secondly it should be considered that the person questioned is not prejudiced and also not motivated to colour the facts. Of course, due allowance should be made for optimism and pessimism.

(iii) ***Schedules and Questionnaires.*** A list of questions regarding the enquiry is prepared and printed. Data are collected in any of the following ways.

(a) *By sending the questionnaire to the persons concerned with a request to answer the questions and return the questionnaire.*

Success in this method depends entirely on the co-operation of the informants. The advantage in this method is that it is less costly, as no enumerators are required and investigations can be completed within a short time.

The disadvantages are—many individual do not return the forms in time and some of the individuals make mistake in filling up the forms.

(b) *By sending the questionnaires through enumerators for helping the informants.*

In this method, enumerators go to the informants to help them in filling the answers. This method is useful for extensive enquiries. It is expensive. *Population census* is conducted by this method. It is essential enumerators should be polite, and have proper training. The implications and scope of each question, to be asked to the informants, should be explained clearly to the enumertors. They should be instructed how to check up apparently wrong replies. They should have intelligence and capacity to cross-examine the informants for finding out the true result.

(iv) ***Local Reports.*** This method does not imply a formal collection of data. Only local agents or correspondents are requested to supply the estimates required. This method gives only approximate results, of course at a low cost.

Questionnaires.

In a statistical enquiry, the necessary information is generally collected in a printed sheet in the form of a questionnaire. This sheet contains a set of questions which the investigator ask to the informant, and the answers are noted down against the respective questions on the sheet. Choice of questions is a very important part of the enquiry whatever be its nature.

For satisfactory investigation a questionnaire should possess the following points :—

(i) The schedule of questions must not be lengthy. Many questions may arise during preparations of questionnaire. If all of them are included, the result is that the persons who are interviewed may feel bored and reluctant to answer all the questions. So only the important questions are to be included.

(ii) It should be simple and clear. The questions should be understandable even by the most uneducated people so that informants do not find any difficulty in furnishing the answers. The factors of simplicity and clarity also imply that the questions should be few so that the informant may not be confused. If possible, the questions should be so set up that require brief answers *viz* 'yes', 'no' or a 'number' etc.

(iii) Each question should be brief and must aim to some particular information necessary for the investigation of the problem. Lengthy questions may be split up into smaller parts, which will be easily grasped by the informants.

(iv) Questions on personal matters like income or property should be avoided as far as possible, as people are generally reluctant to disclose the truth. In such cases, the informations may be collected on guess-work.

(v) The questions should be arranged in a logical sequence. The first part may contain questions like name, age, address etc. and serious or personal questions should be set at the end of the questionnaire so that the informant may answer them when he feels easy with the interviewer.

(vi) The units of informations should be clearly shown in the schedule. For example,

State your age. years.........months.........

what is your weight ? kg.........

Example.

The following form was used in census of population of India 1961, for having a census of Scientific and Technical Personnel.

CENSUS OF INDIA 1961 : SCIENTIFIC & TECHNICAL PERSONNEL

Only a person with a recognised Degree or Diploma in Science, Engineering, Technology or Medicine should fill in this card

READ CAREFULLY BEFORE FILLING IN TICK ( ) WITHIN BRACKETS PROVIDED WHERE APPLICABLE

CENSUS LOCATION CODE

[*Form Contd.*]

1. NAME........................ 2. DATE OF BIRTH....................

3. DESIGNATION & OFFICE ADDRESS....................
(*if employed*)

4. PERMANENT ADDRESS........................

5. (a) Male ( )
(b) Female ( )

6. (a) Never Married ( )
(b) Married ( )

7. On Feb. 1st. 1961 were you

(a) Employed ? ( )
If so, monthly total income
Rs............

(b) Full time student ? ( )

(c) Unemployed if so, how long ? ( )
...........yrs.........mths.

(d) Retired ? ( )

8. ACADEMIC QUESTIONS (ANSWER FULLY)

Degree/Diploma	Subject taken	Division	Year of passing

If employed fill in Qs. 9-12

9. Nature of employment ( )
a. Teaching in School ( )
b. ,, ,, College ( )
c. Technical in Industry ( )
d. ,, Outside Industry ( )
e. Non-Technical ( )

10. Any Research Assignment ( ) ( )
Yes ( ) / No ( )

11. Where employed ?
a. Public Sector ( )
b. Private Sector ( )
c. Self Employment ( )

12. How Employment ?
a. Permanent ( )
b. Temporary ( )
c. On contract ( )
d. Research Scholar ( )
e. Otherwise ( )

Date Signature

Secondary Method.

The main sources from which secondary data are collected are given below—

(i) Official publications by the central and state governments, District Boards,

(ii) Reports of Committees, Commissions,

(iii) Publications by Research Institutions, Universities,

(iv) Economic and Commercial Journals,

(v) Publications of Trade Associations, Chambers of Commerce etc.,

(vi) Market reports, individual research works of Statisticians.

Secondary data are also available from unpublished records of government offices, chambers of commerce, labour bureaus etc.

Editing and Scrutiny.

Secondary data should be used only after careful enquiry and with due criticism. It is advisable not to take them at their face value. Scrutiny is essential because the data might be inaccurate, unsuitable and inadequate. According to Bowley, "It is never safe to take published statistics at their face value without knowing their meanings and limitations........."

Secondary data may, however, be used provided they possess the attributes (*i.e.* qualities) shown below—

1. *Data should be reliable.* The reliability of data depends on the following quiries—

(a) The sources of original collector's informations.

(b) Original compiler's reference.

(c) Method of collection including instructions given to the enumerators.

(d) Period of collections of data.

(e) Degree of accuracy desired and achieved by the compiler.

2. *Data should be suitable.* For the propose of investigation, even the reliable data should be avoided if they are found to be not suitable for the purpose concerned. Data suitable for one enquiry may be unsuitable for the other.

3. *They should be adequate.* Even the reliable and also suitable data may become inadequate sometimes for enquiry. The original data may refer to a certain market price during disturbed period ; for a normal period the above reference will be inadequate.

Representative Data.

Statistical investigation may be complete or partial. Complete investigation, commonly known as *Census*, is the enquiry of each individual item of the universe or population. Partial investigations, commonly known as *Sample Survey*, is the enquiry only to some portions of the universe. The portion selected is called *Sample.*

A sample survey is much less expensive, unlike census. In statistics the word universe (or population) is used to denote a totality of items, covering the whole range of enquiry.

A population is finite or infinite according as it contains finite or infinite number of members. For *example*, the population of class—one cricket players in India is finite while the temperature of Delhi in any day is infinite, for though the temperature varies between two finite limits, it takes up an infinity of values between these two limits. Therefore a sample, as stated is just a portions of population. Selection of sample is an important point. A sample survey would show dependable result only if the sample is a true representative of the universe. The methods of selecting a sample are as follows—

(i) ***Deliberate or Purposive Selection.*** In this method the investigator deliberately selects such items which he feels are the true representatives of the universe. The main drawback of this method is that personal element has a great chance of entering into the selection of Sample.

Two independent purposive sample may give widely different estimates of the same population.

(ii) ***Random Sampling*** (or *chance selection*). It is the method of drawing a sample from an universe, such that every member of the universe has an equal chance of being including in the sample. It is in fact a lottery method of selection.

Example. To select a random sample of 4 boys from 99 boys, other than applying lottery method, we can select the sample of 4 boys by the following way. Let the roll nos. of the boys are 1, 2, 3,......, 98, 99.

Tables of random numbers from digits 1, 2, 3,...9, 0 arranged in rows and columns in purely *haphazard manner*, are shown below—

2	5	7	9	0	5	6	7	1	3
6	7	2	1	5	2	0	1	2	1
3	4	1	3	2	7	9	0	5	4

Let us now take any row and column, make 4 two-digit figures successively as they occur. Thus if we start from second row and first column and move horizontally, we will find numbers, 67, 21, 52, 01. So the boys having the above roll numbers will be the required random samples. If a particular number occurs twice or more, then only one number is to be retained.

When the population is homogeneous in respect of a particular characteristic, random samples yeild better results in respect to other type of sample.

(iii) ***Stratified Sampling.*** Under this method, the population is *purposively* subdivided into several parts (known as *Strata*), then sample from each stratum is choosen at *random*. It may be noted, subdivision of population is *purposive*, while choosing of sample is purely *random*.

So this is a *mixed sampling* of purposive and random sampling. Stratified sampling is used when the population is heterogeneous.

(iv) ***Systematic Sampling.*** In this method every r-th member of the population arranged serially is drawn, of course first member is to be selected at random, from first r-members of the populations.

Example. Suppose we are to select a systematic sample of 5 boys out of 150 boys, numbered serially. Here the number of populations 30 times the sample size. Any number between 1 to 30, say 21, is selected at random. Now every subsequent 30th number i.e. 51, 81, 111, 141 are included in the sample. Required sample of 5 boys will be of numbers 21, 51, 81, 111, 141.

Law of Statistical Regularity.

According to the rules of the theory of probability, if from the universe a moderately large sized sample is chosen at *random*, it is likely that *on an average* the sample chosen will have the same characteristics as the universe. In statistics, this law is known as *Law of Statistical Regularity.* The theory of probability tells us of the mathematical expectation of happening or failure of an event, and on this basis the law of statistical regularity tells us that random selection from the universe is very likely to give a representative sample.

It may be noted that *any* number of samples will not yield exactly the same results as a study of universe would. The same probability of error diminishes with the increase in number of items taken in the sample. That is, the larger the sample, the more reliable are the results.

Law of Inertia of Large Numbers.

It is a corollary of the above law. We know the larger the sample greater would be the accuracy, for in large numbers the chances of compensatory are more. The production of rice in Burdwan district might vary at large, year to year, but in West Bengal State the narration of the same productions would be less. In the same way productions of the same commodity in India would show still less variations. This phenomenon is generalised as *Law of Inertia of large numbers,* which implies the large numbers are relatively more stable than small ones.

But it does not mean that the property of inertia does not allow any change with passage of time. It signifies that large numbers are more constant and stable than small ones. Above all there is no striking change in large numbers.

Statistical Errors.

The word *error* is used in a special sense in statistics. It is the difference between the true value and the estimated value of a quantity. It shows by how much an estimate of a measurement falls short of or exceeds the true measurement. It does not mean the same thing as mistake. In statistics mistake means a wrong calculation or wrong method used in collection or analysis.

Sources of Errors.

(1) *Errors of origin*—errors resulting from inappropiate or faulty definition of units.

(2) *Errors of inadequacy*—errors due to inadequate size of sample or incomplete information.

(3) *Errors of manipulation*—errors due to manipulation in counting, measuring, weighing or approximation.

Types of Errors.

There are two types of errors (a) *Absolute error* and (b) *Relative error.*

Absolute error is the difference between the true value and the estimate of a quantity, while relative error is the ratio of absolute error to the estimate.

If again the relative error is expressed as percentage, then it will be a *percentage error*. For example, if true value (price) of a quantity is Rs. 101, estimate value is Rs. 100, then

Absolute error = 101 − 100 = Re. 1, relative error = $\frac{1}{100}$ = ·01 and percentage error = ·01 × 100 = 1%.

Illustration. (A Specimen of a Blank Form).

The form recently used by the Government of West Bengal for Census of Employees of State Government, Local Bodies etc. is shown below :

Form No.S-5-7 **State Govt. Office/**
Non-Govt. Office

GOVERNMENT OF WEST BENGAL

BUREAU OF APPLIED ECONOMICS & STATISTICS

Census of Employees of State Government, Local Bodies, State Public Sector Undertakings and Autonomous Bodies

As on..........................

INDIVIDUAL SLIP

Name of Employee..
Name of Office...
Address (in full)...
Department (for Govt. Offices only).......................................
Branch/Directorate (for Govt. Offices only)..............................
District..........................P.S......................Block..................
Town/Village...................Sector...

1. *Sex (code).................□ 2. Date of birth..........................
3. Date of entry in service...... 4. *Caste (code).....................□
5. Designation...
6. Name of service...
7. Scale of Pay....................... 8. *Status (code)....................□

(*Form Contd.*)

9. Emolument (in round rupees for March, 197) :

(a) Basic Pay.................. (b) Dearness Pay.........................

(c) Ad-hoc Pay................ (d) Special Pay..........................

(e) Personal Pay............. (f) Dearness Allowance.................

(g) Interim Dearness Allowance............... (h) Medical Allowance.................

(i) House Rent Allowance............... (j) Compensatory Allowance..........

(k) Other emoluments : (specify) (i) ....

(ii) ....

(iii) ....

(iv) Additional Dearness Allowance......

(l) Total gross emolument..

(m) Total deduction from emolument (P.F., Recovery of Advance, etc.).............................

(n) Net emolument after deduction.......................................

10. Place of posting :............ (a) *District (code).....................□

(b) *Rural/Urban (code).............□

11. Date of present posting...

12. Place of residence :......... (a) Type : Own House/Government Quarter/Rented House/Rent-free Quarter.

(b) Address :..................................

13. *Educational qualification (code)□...................................□

14. Total number of members in the family.............................□

15. Total number of Government employees in the family............□

Signature of Employee..........................Date.....................

Signature of Enumerator........................Date.....................

CODE

Serial No. 1—Sex :	(i) Male — 1,		(ii) Female —	2
Serial No. 4—Caste :	(i) Scheduled Cast		—	1
	(ii) Scheduled Tribe		—	2
	(iii) Others		—	3

Serial No. 8—Status :	(i)	Permanent	—	1
	(ii)	Permanent Status	—	2
	(iii)	Quasi Permanent	—	3
	(iv)	Temporary	—	4
	(v)	Part-time	—	5
	(vi)	Piece-rate	—	6
	(vii)	Contingency Menial	—	7
	(viii)	Work-charged	—	8
	(ix)	Contract	—	9
Serial No. 10 (a)—District :	(1)	Calcutta	—	01
	(2)	Burdwan	—	02
	(3)	Birbhum	—	03
	(4)	Bankura	—	04
	(5)	Midnapore	—	05
	(6)	Howrah	—	06
	(7)	Hooghly	—	07
	(8)	24-Parganas	—	08
	(9)	Nadia	—	09
	(10)	Murshidabad	—	10
	(11)	West Dinajpur	—	11
	(12)	Malda	—	12
	(13)	Jalpaiguri	—	13
	(14)	Darjeeling	—	14
	(15)	Cooch Behar	—	15
	(16)	Purulia	—	16
Serial No. 10 (b)—Rural/Urban :	(i)	Rural	—	1
	(ii)	Urban	—	2
Serial No. 13—Educational qualifications :	(1)	Below School Final Standard	—	01
	(2)	School Final or equivalent	—	02
	(3)	Higher Secondary or equivalent	—	03
	(4)	Intermediate/Twelve Class pass or equivalent	—	04
	(5)	Technical Diploma	—	05
	(6)	Graduate (Arts, Science, Commerce)	—	06
	(7)	Graduate (Engineering)	—	07
	(8)	Graduate (Medical—M.B,B.S.)	—	08
	(9)	Post-Graduate (Arts, Science, Commerce)	—	09
	(10)	Post-Graduate (Engineering)	—	10
	(11)	Post-Graduate (Medical)	—	11

EXERCISE 1

1. Distinguish between Primary data and Secondary data. State various methods of collecting primary data and comment on their relative advantages. [I. C. W. A. Jan. 1972]

2. Define Secondary data. State their chief sources and point out the dangers involved in their use and what precautions are necessary before using them. [C. A. Nov. 1967]

3. Define statistical unit, mention the usual kinds of units employed in statistical work. What are the essential points to be observed in the choice of a good unit ?

Giving appropiate reasons, state what units can be used for the following cases—

(i) Production of cotton textile industry,

(ii) Labour employed in the industry,

(iii) Consumption of electricity. [C. A. May 1967]

4. What are statistical units ? How would you define them ? Describe the various types of statistical units and explain.

5. Distinguish between census and sample method of investigation. What are their relative merits and defects.

6. Distinguish between Random and Stratified Sample.

7. What are the essentials of a good questionnaire ? A certain state has just passed an enactment making attendance at school compulsory for all children between ages 5-15. You are asked to collect all statistics that might be necessary for the purpose of enforcing the Act.

State how would you proceed with the work and what statistics you would collect. Draw up a suitable questionnaire blank form to collect necessary information.

8. It is required to collect information on the economic conditions of textile mill workers in Bombay. Suggest a suitable method for collection of Primary data. Draft a suitable questionnaire of about ten questions for collecting this information. Also suggest how you will proceed to carry out statistical analysis of the information collected, [I. C. W. A. June 1976]

3

CLASSIFICATION AND TABULATION

Introduction.

The data collected or compiled by the methods discussed in the previous chapter are usually voluminous, crude in form and are known as *raw* data. They are not directly fit for any statistical purpose. For the purpose of any analysis and interpretation, the data require proper arrangements and modifications.

Classification.

It is the process of arranging data into different classes or groups according to resemblances and similarities. An ideal classification should be unambiguous, stable and flexible.

Object.

The objects of classification are many. It clearly shows the points of similarity, dissimilarity. It prepares the ground for comparisons and analysis by orderly arrangements of data.

Types of Classification.

There are two types of classification depending upon the nature of data.

(i) classification according to attribute—if the data is of a descriptive nature having several qualifications *i.e.*, males, females, literate, illiterate etc.

(ii) classification according to class-intervals—if the data are expressed in numerical quantities, *i.e.*, ages of persons vary and so do there heights and weights.

Classification according to Attributes.

(i) *Simple classification* is that when only one attribute is present *i.e.*, classification of persons according to sex—males or females.

(ii) *Manifold classification* is that when more than one attribute are persent simultaneously *i.e.*, classification of persons regarding deafness sex-wise. Now we find that there are two attributes—deafness and sex. A person may be either deaf or not deaf, further a person may be a male or female. The data, thus, are to be divided into four classes (a) males who are deaf (b) males who are not deaf (c) females who are deaf (d) females who are not deaf. The study can be further continued, if we find another attribute say religion.

Classification according to Class-intervals.

This type arises when direct measurements of data is possible. Data relating to height, weight, production etc. come under this category. For instance, persons having weight say 100—110 lbs. can form one group, 110—120 lbs. another group and so on. In this way data are divided into different classes; each of which is known as *class-interval*. Number of items which fall in any class-interval is known *class-frequency*. In the class-intervals mentioned above, the first-figures in each of them are the *lower limits*, while the second figures are the *upper limits*. The difference between the limits of a class-interval is known as *magnitude* of the class-interval. If for each class-intervals the frequencies given are aggregates of the preceeding frequencies, they are known as *cumulative frequencies*. The frequencies may be cumulated either from top or from below.

In general, the class-intervals should be of equal magnitude. If the size of the class-interval is unequal it may give a misleading impression, and in such cases, comparison of one class with the other may not be possible.

Methods of forming Class-intervals.

The class-intervals *i.e.* 100—110, 110—120, 120—130 etc. are *Overlapping*. Difficulty arises when placing an item, say, 110 in the above class-interval. Whether 110 lbs. should be placed in the class-intervals 100—110 or 110—120. Now in this method, known as *Exclusive method*, an item which is identical to the upper limit of a class-interval is *excluded* from that class-interval, and is included in the next class-interval. So the item 110 lbs. will belong to the class interval 110—120. For all practical uses, 100—110 means 100 and less than 110, again 110—120 means 110 and less then 120, and so on.

Again the class-intervals may be formed as 100—109, 110—119, 120—129 etc. In this method, known as *Inclusive method*, also difficulty arises when there is an item lying between the upper limit of a class and lower limit of the next class. The above class-intervals may also be arranged as 100—109·5, 110—119·5 and so on.

Now it shows whatever be the upper limit in the first class *i.e.* 110, 109, 109·5 or 109·9, it is always less than 110.

It may be noted magnitude in every case is 10.

15339

Class-intervals with Cumulative Frequencies.

If the class-frequencies are given as cumulative class-frequencies, then the class-intervals also are expressed only by their upper limits preceded by the word 'below' (or less than) or 'above' (or more than) according as the frequencies are cumulated from the top or bottom. Before treating with such data for any statistical purpose, it is necessary to convert it into usual class-intervals with their corresponding class-frequencies. From the following *example*, the idea of converting the cumulative frequencies to usual frequencies will be clear.

(a) *class-frequencies cumulated from top*

	Weights (lb.)	Persons
Below	110	10
„	120	15
„	130	17
„	140	21
„	150	27

(b) *class-frequencies cumulated from bottom*

	Weights (lb.)	Persons
Above	100	27
„	110	17
„	120	12
„	130	10
„	140	6

Now the usual type of class-intervals having class-frequencies will be as follows—

Weights (lb.)	Persons
100—110	10
110—120	5
120—130	2
130—140	4
140—150	6

Statistical Series.

If things or attributes are measured, counted or weighted, and they are placed one after another, the result is a statistical series. In brief, a *statistical series* may be defined as things or attributes arranged in some logical or systematic order.

Types.

There are three bases of classifications of data : *time, space* and *condition* and consequently there are three types of statistical series : (i) *time* or historical (ii) *spatial* (iii) *condition.* In the first if the data collected relates to past or present. Now if the figures of enrolment of students in a college for the last ten years are arranged in order, they would form a time (or historical) series. In the second, if the data collected change in relation of place (and not in relation to time), the series is known as spatial series. Production of wheat in India for a particular year, for different states, would form a spatial series. In the third, data are recorded on the basis of physical condition. If heights, weights, or ages of 100 students are recorded, the different data, arranged in order, shall constitute a condition series.

Discrete and Continuous Series.

Statistical series may be either discrete or continuous. A discrete series is formed from items which are exactly measurable. Every unit of data is separate, complete and not capable of divisions. For instance, the number of students obtaining marks exactly 10, 14, 18, 20, can easily be counted. But phenomenon like height or weight cannot be measured exactly or with absolute accuracy. So the number of students (or individuals) having height exactly 5′2″ cannot be counted. Exact height may be either side of 5′2″ by a hundredth part of an inch. In such cases, we are to count the number of students whose heights lie between 5′0″ to 5′2″. Such series are known as 'continuous series.

Example.

Discrete Series		Continuous Series	
Marks	No. of Students	Height (inch)	No. of Students
10	12	58—60	6
14	16	60—62	10
18	15	62—64	13
20	7	64—66	11

Tabulation.

Tabulation is a systematic and scientific presentation of data in a suitable form for analysis and interpretation.

After the data have been collected, they are tabulated *i.e.* put in a tabular form of columns and rows. The function of tabulation is to arrange the classified data in an orderly manner suitable for analysis and interpretation. Tabulation is the last stage in collection and compilation of data, and is a kind of stepping-stone to the analysis and interpretation.

A table broadly consists of five parts :—

(i) *Number and Title* indicating the serial number of the table and the subject matter of the table.

(ii) *Stub* i.e. the column indicating the headings of rows.

(iii) *Caption* i.e. the headings of the column (other than stub).

(iv) *Body* i.e. figures to be entered in the table.

(v) *Foot Note* is source from which the data have been obtained.

Thus a table should be arranged as follows—

Table No.
Title

	Caption	Total
Stub	Body	
Total		

Types of Tabulation.

Mainly there are two types of table—*Simple* and *Complex.* Simple tabulation reveals information regarding one or more groups of independent question, while complex table gives information about one or more inter-related questions.

One way table is one that answers one or more independent questions. So it is a simple tabulation. The following table will explain the point—

Table 1. Daily wages in Rs. obtained by 50 workers in a factory.

Wages (Rs.)	*No. of Workers*
4—6	20
6—8	[illegible]
8—10	10
10—12	7
12—14	4
Total	50

The table shows the number of workers belonging to each class-interval of wages. We can now easily say that there are 20 workers, obtain wages between 4 and 6 (the minimum range) and there are 4 workers, obtain wages between 12 and 14 (the maximum range). So this table reveals information regarding only one characteristic of data *i.e.* wages of workers.

Two-way table shows sub-division of a total and is able of answering two mutually dependent questions. In the above table (no. 1), if the workers are divided into sex-wise, then we would get a two-way table as follows—

Table 2. Daily wages in Rs. obtained by 50 workers (sex-wise)

Wages (Rs.)	*No. of Workers*		
	male	*female*	*total*
4—6	12	8	20
6—8	6	3	9
8—10	6	4	10
10—12	4	3	7
12—14	4	0	4
Total	32	18	50

The above table shows the wages obtained by workers and sex-wise distribution of workers in question.

Three way table sub-divides a total into three distinct catagories and is capable of answering three mutually dependent questions. In the above table (no. 2), if the workers are divided into resident and non-resident (in the factory area), we would get a three way table as given below—

Table 3. Wages (Rs.) obtained by 50 workers in a factory (sex-wise and resident-wise)

wages (Rs.)	*Male*			*Female*			*Total*		
	resident	*non resident*	*total*	*resident*	*non resident*	*total*	*resident*	*non resident*	*total*
4—6	4	8	12	4	4	8	8	12	20
6—8	3	3	6	2	1	3	5	4	9
8—10	3	3	6	2	2	4	5	5	10
10—12	4	0	4	2	1	3	6	1	7
12—14	3	1	4	0	0	0	3	1	4
Total	17	15	32	10	8	18	27	23	50

The table, shown above, gives the informations about (i) wages (in Rs.) obtained by workers (ii) sex-wise distribution of these workers and (iii) distribution of workers on the basis of residence.

If the table is again classified on the basis of different religions, states, nationalities, etc. it will give an example of *Manifold tabulation.*

Rules and Precautions for Tabulation.

It is necessary to put down certain rules and precautions in drawing up the tables. For the construction of tables, the following points should be observed—

(1) If the data are too large, then several separate tables should be drawn instead of a single table. In such case, a single table will confuse eye and may lead to great difficulty in following the columns and rows at a glance. Of course, each table should be complete in itself and should serve a particular purpose.

(2) The table should suit the size of the paper on which it is drawn. So the width of columns and rows should be decided properly. Totals, averages, percentages and the numbers for comparisons, should be placed close together as far as possible. Unimportant data may be placed in miscellaneous group.

(3) For separating data of one class from that of another class, thick lines should be used. Of course, thin lines may be used for separating the sub-division classes.

(4) The table should be given a suitable title. The title or titles of sub-headings (*i.e.* captions and stubs) should be self-explanatory. The column headings (i.e. captions) should indicate the unit used *i.e.* height in inches, price in rupees, weight in pounds etc.

(5) Large digits may be approximated to thousands or, lakhs etc. This would reduce the unnecessary details.

(6) Explanatory notes should be given always as footnotes and must be complete in itself.

The *source* from which the data is obtained should be indicated in the footnote. This will help the reliability of the data. In case of any discrepancy or inconsistency found in the data, attention must be drawn by using footnote using reference like $_{1},_{2},_{3}$ or*,† etc.

(7) The items in the table should be arranged with some logical order. They may, however, be arranged in order of magnitude or alphabetical, geographical or in other suitable manner.

(8) Before entering in the table, items should be checked up carefully. Arrangement should also be made for cross-checking. Over-writing in the table should be avoided.

Example 1.

Construct a blank table in which could be shown at different dates and in five industries the average wages of the four groups, males and females, eighteen years and over, and under eighteen years.

(C. U. M. A. 1963)

Average Wages of Employees in 5 Industries

Industry	As on..............(date)						As on..............(date)					
	Under 18 yrs.			18 yrs. & over			Under 18 yrs.			18 yrs. & over		
	male	*female*	*total*	*male*	*female*	*total*	*male*	*female*	*total*	*male*	*female*	*total*
1												
2												
3												
4												
5												
Total												

Example 2.

Prepare a neat table, paying attention to headings, double lines spacing etc. showing fully the information in the following report as clearly as possible—

During the quinquennium 1935—39, there were in Great Britain 1,775 cases of industrial diseases made up of 677 cases of lead-poisoning, 111 of other poisoning, 144 of anthrax and 843 of gassing. The number of deaths reported was 20 p.c. of the cases for all the four diseases taken together, that for lead poisoning was 135, for other poisoning 25, and that for anthrax was 30.

During the next quinquennium 1940—44, the total number of cases reported was 2807 higher. But lead-poisoning cases reported fell by 351 and anthrax cases by 35, other poisoning cases increased by 748 between two periods. The number of deaths reported decreased by 45 for lead-poisoning but decreased only by 2 for anthrax from the pre-war to post war quinquennium. In the latter period 52 deaths were poisoning. The total number of deaths reported in 1940—44 including those from gassing was 64 greater than, in 1935—39.

Industrial diseases in Great Britain

Quinquennium	Cases of					Deaths					Remarks
	Lead Poisoning	Other Poisoning	Anthrax	Gassing	Total	Lead Poisoning	Other Poisoning	Anthrax	Gassing	Total	
1935—39	677	111	144	843	1775	135	25	30	165	355	
1940—44	326	859	109	3288	4582	90	52	28	249	419	

Mechanical Aid to Computation

Mechanisation become more and more useful in the field of computation of any practical problem. If the problem is a small one we can do it by manual, but if the problem is large or complicated we use to take mechanical aid of computation. Mechanical aids to computation include machine types like.

(1) Listing and adding machines.

(2) Calculating machines.

(3) Accounting machines.

(4) Tabulating and Statistical machines.

Adding machines are generally either hand-operated or electrically operated. Ordinary adding machines can only do addition process, when machine print the data in a tape in addition to adding, is called adding and listing machine.

Calculating machines can do multiplication and division in addition to adding. They also do subtraction functions. These machines are mostly operated manually or electrically.

Accounting machines can do calculations and post the result of calculation in a card at an appropriate place. Such as cash register, cost sheet preparation, stores ledgers, journal books. Material issue analysis etc. are done by this type of machine.

Tabulating and Statistical machines are punch card processing machines. Here all basic informations are punched in a card and the cards are fed into machines which read the instructions and process the fed data and give the result in a form of printed out. The essential function of any punched card machines are as follows :

(i) Card punch.

(ii) Card punch verifier.

(iii) Interpreter.

(iv) Sorter.

(v) Collator.

(vi) Reproducer.

(vii) Calculator.

(viii) Tabulator.

Computer is an improved type of tabulating and statistical machines. Two types of computers are generally used in computations such as (a) Analog Computer and (b) Digital Computers. Analog Computers are required to measure physical quantity as energy, velocity etc. Digital computers are used in quantitative analysis. It can do all sorts of numerical calculation if proper instruction has been given to it.

EXERCISE 2

1. Define classification, what part does it play in Statistics ? State the different methods of classification of Statistical data.

2. Discuss the function and importance of tabulation in a scheme of statistical investigation. What precautions should be taken in tabulation of data ?

3. What is a statistical table ? Explain clearly the essentials of a good table.

Prepare a blank table showing the particulars relating to students of the Bombay University classified according to their ages, sex, faculty and three important religions. [C. A. Nov. 1969]

4. Define a statistical table and state the essentials of a good table.

5. Draw a blank table showing exports and imports during the years 1960, 1961, 1962, 1963, 1964 relating to Ports Bombay, Calcutta, Madras and other Ports. The table should provide of the values and the balance of trade and the totals for each year. [C. A. Nov. 1968]

6. (a) Discuss the general rules which should be followed in tabulating statistical data and state the various types of tabulations and their uses.

(b) Draw up a blank table with a suitable title, headings, double lines etc. in which could be shown the population of India by states during the 1961 census, classified according to sex, age groups and the two principal livelihood categories—Agricultural and non-agricultural. [I. C. W. A. Jan. 1968]

7. The total number of accidents in Southern Railway in 1960 was 3500 and it decreased by 300 in 1961 and by 700 in 1962. The total number of accidents in metre gauge section showed a progressive increase from 1960 to 1962. It was 245 in 1960 ; 346 in 1961 ; and 428 in 1962. In the metre gauge section, "Not Compensated" cases were 49 in 1960, 77 in 1961, and 108 in 1962. "Compensated" cases in the broad gauge section were 2867, 2587 and 2152 in these three years respectively.

From the above report, you are required to prepare a neat table as par the rules of tabulation. [C. A. Nov. 1971]

8. (a) What are the different parts of a statistical table ?

(b) Present the following information in a concise tabulator form and indicate which type of lamp shows the greatest wastage during manufacture :

"Lamps are rejected at several manufacturing stages for different faults. 12,000 glass tubes are supplied to make 40-watt, 60-watt and 100-watt lamps in the ratio 1 : 2 : 3. At the stage I, 10% of 40-watt, 4% of the 60-watt, and 5% of the 100-watt bulbs are broken. At the stage II, about 1% of the remainder lamps have broken filaments. At the stage III, 100-watt lamps have badly soldered caps and half as many have crooked caps ; twice as many 40-watt and 60-watt lamps have these faults. At the stage IV, about 3% are rejected for bad type making and 1 in every 100 are broken in the packing which follows." [I. C. W. A. July 1971]

9. Draw up a blank table in which could be shown the number of persons employed in six industries on two different dates, distinguishing males from females and among the latter, singles, married and widows. [I. C. W. A. Jan. 1973]

10. In the Sutton coalfield region the number of cinema admissions in the four quarters of 1950 was ('000) : 11,008 ; 9,998 ; 9,933 and 9,406. For the four quarters of 1951 they amounted to ('000) : 10, 521 ; 9,677 ; 9,369 and 9,568. The corresponding number of television licences per 1000 population was 12 ; 18 ; 24 ; 37 ; 52 ; 64 ; 69 and 81. In the Holme Moss region cinema admissions during the eight quarters of the two years were ('000) ; 18,290 ; 16,420 ; 16,973 ; 15,937 ; 17,940 ; 16,431 ; 16,336 and 16,136. The quarterly number of television licences per 1000 of the population in the Holme Moss region during 1950 and 1951 was 2, 3, 4, 6, 9, 13, 14 and 29. Transmission of television started in the Holme Moss region during the fourth quarter of 1951.

Arrange the above information in a tabular form. What do you deduce from these figures ?

11. Prepare a blank table to show the exports of three companies A,B,C to the five countries. U.K., U.S.A., U.S.S.R., France and West Germany, in each of the years 1970—74.

[I.C.W.A. Dec. 1975]

12. What is the purpose of tabulation of statistical data ? What general rules should be observed in constructing a statistical table ? [I. C. W. A. Dec. 1974]

13. Draw up a blank table to show the numbers of candidates, sex-wise, appearing for the Pre-University, First Year, Second Year and Third Year examinations of a University in the faculties of Arts, Science and Commerce in a certain year. [I. C. W. A. Dec. 1974]

14. State briefly the requirements of a good statistical table.

Prepare a blank table to show the distribution of population of the various States and Union Territories of India, according to sex and literacy. [I. C. W. A. June, 76]

15. Represent the following information in suitable tabular form with proper rulings and headings :

The annual report of the Ishapur Public Library reveals the following points regarding the reading-habits of its members.

Out of total 3,713 books issued to the members in the month of June 1970, 2,100 were fictions. There were 467 members of the library during the period and they were classified into five classes A, B, C, D and E. The number of members belonging to the first four classes were respectively 15, 176, 98 and 129 ; and the number of fictions issued to them were 103, 1187, 647 and 58 respectively. Number of books, other than textbooks and fictions, issued to these four classes of members were respectively 4, 390, 217 and 341. Textbooks were issued only to members belonging to the classes C, D and E and the number of textbooks issued to them were respectively 3,317 and 160.

During the same period, 1246 periodicals were issued. These included 396 technical journals of which 36 were issued to members of class B, 45 to class D and 315 to class E.

To members of the classes B, C, D and E the number of other journals issued were 419, 26, 231 and 99 respectively.

The report, however, showed an increase by 3·9% in the number of books issued over last month, though there was a corresponding decrease by 6·1% in the number of periodicals and journals issued to members. [I. C. W. A. June, 77]

16. Represent the following data in a tabular form :

A firm processes a certain raw materials by the use of two major types of equipments, called stills and retorts. Four different production processes are available to the firm. One unit of each of the processes I, II, III and IV will weekly treat 100 tones of raw meterial. Processes I and II will absorb 7 per cent and 5 per cent of the weekly capacity of the firm's stills and 3 per cent and 5 per cent respectively of the weekly capacity of the retorts. Processes III and IV will absorb 3 per cent and 2 per cent of the weekly capacities of the firm's stills and 10 per cent and 15 per cent respectively the weekly capacity of the retorts. The four processes I, II, III and IV will respectively yield a final product worth $ 1100, $ 1120, $ 1130 and $ 1150 but will consume raw material of value $ 1000, $ 1000, $ 1000 and $ 1000 respectively. The other direct costs required by the four processes are worth $ 50, $ 60, $ 40 and $ 60 respectively.

[I. C. W. A. June, 79]

17. Present the following data in a tabular form :

A certain manufacturer produces three different products 1, 2 and 3. The product 1 can be manufactured in one of the three plants : *A*, *B* and *C*, However the product 2 can be manufactured in either plant *B* or plant *C*, whereas plant *A* or *B* can manufacture product 3. The plant *A* can manufacture per hour 10 pieces of 1 or 20 pieces of 3. 20 pieces of 2, 15 pieces of 1 or 16 pieces of 3 can be manufactured per hour in plant *B*.

Whereas *C* can produce 20 pieces of 1 or 18 pieces of 2 per hour. Wage rates per hour are Rs. 1·50 at *A*, Rs. 3·00 at *B* and Rs. 2·00 at *C*. The cost of running plants *A*, *B*, *C* are respectively Rs. 200/-, Rs. 100/-, Rs. 250/- per hour. The materials and other costs directly related to the production of one piece of the products are respectively Rs. 10 for 1, Rs. 12 for 2 and Rs. 15 for 3. The company plans to market the product 1 for Rs. 15 per piece, the product 2 for Rs. 18 per piece and the product 3 for Rs. 20 per piece. [I. C. W. A. Dec. 79]

4

PRESENTATION OF DATA, GRAPHS AND CHARTS

Introduction.

In the previous chapters we have discussed how huge statistical data are condensed and presented in an intelligent form of a table. Various statistical methods like classifications, tabulations, averages and index numbers etc. reduce the complexity of statistical data in a simpler form. Now classification and tabulation are meant for systematic presentation of data, while measures of central tendency and index numbers (these chapters will be discussed later on) help for comparing data by converting into single figures. These chapters have their own limitations. One more effective method of representing data is by the help of graphs, charts and diagrams, by means of which the true significance of a set of figure can be easily grasped. Of course, it is true that graphs and charts add nothing to the information already obtained but they bring out clearly the relative importance of different figures, and often, are necessary in finding out the trend of the values or variations in the values, in relations to time. The special feature of diagrams and graphs is that they present dry and uninteresting statistical facts in the shape of attractive and appealing pictures and charts.

Usefulness.

The advantage in diagrammatic presentation of data is that diagrams and charts are attractive to common people. Common people in general, avoid figures, but always search for pictures and diagrams, even when reading general books. Graphical representations are particularly useful when comparisons are to be made between two or more sets of data. To an economist, difficult theories can be easily understood if proper diagrams are used. Diagrams save much valuable time, which would be lost in grasping the significance of numerical data.

Limitations.

The charts do not show details, which is possible in a table. Graphical representation reveals only the approximate position, where as in a table, we can show exact figures. If a statistician or an investigator wants to do an exhaustive study of figures, then the utility of diagrams is not much. Representation of data by means of graphs or diagrams may often give misleading impressions to people. Advertisers or politicians misuse this form of representations of facts and try to mislead the common people.

Functions.

The following two functions are served by graphs and diagrams :

(1) They make complex data simple and easily understandable.

(2) They help to compare the related data, placing the graphic or diagrammatic representation near to each other.

GRAPHIC REPRESENTATION

Graphs are useful for representing data relating to time or for representing frequency distribution.

Construction.

Two straight lines are drawn cutting each other at right angles. The horizontal line (XX′) is called *abscissa* or X-axis, while the vertical line (YY′) is known as *ordinate* or Y-axis. The point (O) at which the lines meet is known as *origin*. (See the figure).

Distances measured from O towards right or above are reckoned as positive while those measured towards the left or downwards as negative. Thus XX′ and YY′ divide the plane *i.e.*, graph paper into four parts, known as *Quadrants*. All points in the plane are located by two co-ordinates drawn parallel to the axes.

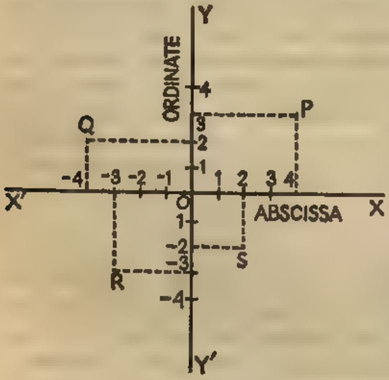


Fig. 1

For each axis, a convenient scale is chosen. It is not necessary that the two scales of the axes should be same. In the above graph, scales of the axes are same and the following points have been plotted.

Points	X	Y
P	4	3
Q	−4	2
R	−3	−3
S	2	−2

The dotted lines forming P, Q, R and S need not be shown.

Only the points should be shown. Thus we find, that a point on a graph paper is a function of two variables.

A graph must be accompanied by its *heading* showing in detail the nature of the graph.

GRAPHS OF TIME SERIES

Natural Scale.

Graphs of continuous time series are known as *Historigram*, which may be constructed on natural scale or on ratio scale. First on natural scale, we will discuss the following graphs.

1. *Absolute Historigram of one variable*—changes of a single variable over a period of time.

2. *Absolute Historigram of two or more variables*—changes of two or more variables over a period of time.

3. *Index Historigram*—If the values are represented by index numbers and if these indices, instead of actual values, are plotted then Index Historigram is obtained.

(1) Absolute Historigram (*or Line Chart or Linear Graph*).

Let x and y are two variables (discrete or continuous) such that for each value of x, there corresponds a value of y. If for one value of x along X-axis, the corresponding value of y along Y-axis be plotted then corresponding to a set of values of x, we get a set of values of y. Then the straight lines or curve obtained by joining the corresponding points is known as Absolute Historigram.

Example. The monthly productions of bi-cycles in a certain factory are as follows—

Jan.—70, Feb.—90, March—80, April—120, May—100, June—120, July—110, Aug.—125, Sept.—130, Oct.—150, Nov.—100.

—Draw a Historigram (absolute) to represent the above data.

Since the values of both the variables are positive we shall draw only one quadrant (*i.e.*, 1st quadrant) is which both the variables are positive.

We represent the months along X-axis and corresponding productions along Y-axis according to the scale mentioned below—

1 division along X-axis = 1 month, 1 division along Y-axis = 10 units.

Graph showing monthly productions of Bi-cycles

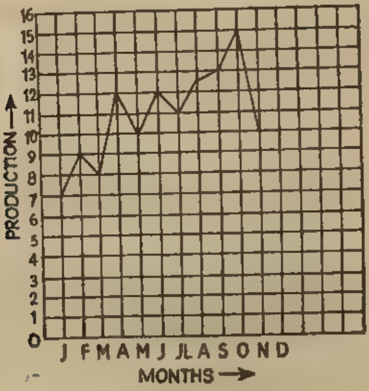


Fig. 2

Use of False Base Line.

If size of items is big and the vertical scale starts from zero, the curve would be almost on the top of the graph paper, as shown above. In order to use utmost space of the graph and to be technically correct also, the false base line is used.

Generally the vertical scale is broken into two parts and some blank space is left in between them. The lower part starts from zero and the upper part starts with a value or nearly equal to the minimum value of the variable. Usually saw-tooth lines are used to break the vertical scale. The false base line is used to represent the above example graphically.

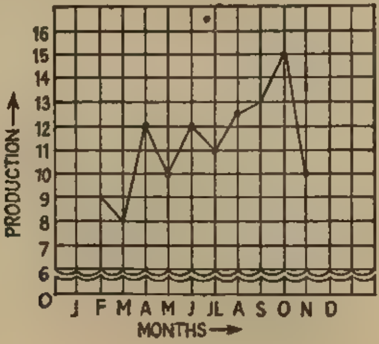


Fig. 3

(2) Absolute Historigram of Two or More Variables.

Two or more variables (of same unit) can also be drawn on the same graph paper. The procedure of drawing is the same as in the previous case. In such case, we will find two or more curves.

Example. The table below gives the income, expenditure and profit (or loss) of a shop during the whole year :

Months	*Income* (*Rs.*)	*Expenditure* (*Rs.*)	*Profit* (*Rs.*)
Jan.	700	500	200
Feb.	800	550	250
March	500	600	–100
April	550	700	–150
May	600	500	100
June	550	400	150
July	700	500	200
Aug.	800	550	250
Sept.	750	400	350
Oct.	700	500	200
Nov.	800	550	250
Dec.	750	450	300

Draw income, expenditure and profit graphically on the same graph paper.

Scale : 1 division along X-axis = 1 month
1 " " Y-axis = Rs. 100 (profit)
1 " " Y′-axis = Rs. 100 (loss)

Graphs of Income, Expenditure and Profit

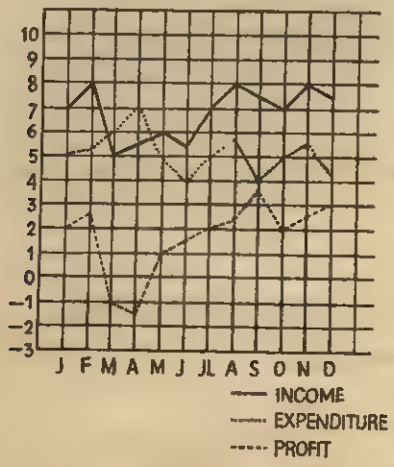


Fig. 4

Note. The false base line cannot be used in this type of graph, where profit and loss are to be shown.

(3) Index Historigram.

The drawing is similar to the previous graph, but only the index numbers are to be plotted instead of variables.

Example.

Small Savings during the year 1960-1964 are shown below, including the index numbers. Draw the index historigram.

Year	*Savings (Rs.)*	*Index* (1961 = 100)
1960	250	125
1961	200	100
1962	300	150
1963	316	158
1964	400	200

Index Historigram of Small Savings (1960-1964), (1961 = 100)

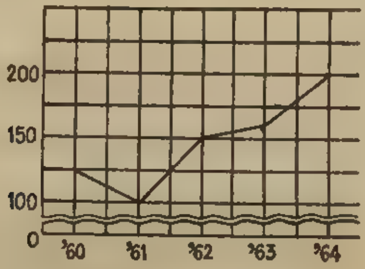


Fig. 5

Note. Index Historigram may be also of two or more variables. The drawing is similar to the drawing of Absolute Historigram of two or more variables.

OTHER GRAPHS

There are certain other graphs, which are becoming popular.

(1) Mixed Graphs.

The graphs are prepared to study the inter-related values. In such graphs one variable is usually shown by bar-diagram and the other by a curve.

Example.

Quantity and Price of a Commodity (1965—70) are shown below. To draw a mixed graph.

Year	1965	'66	'67	'68	'69	'70
Quantity (Kg.)	55	40	35	50	40	25
Price (Rs.)	250	190	180	240	227	140

The above figures can be represented by means of a mixed graph. The quantity will be represented by vertical bars (the length of the bars is proportional to the values they represent) and the price will be shown by historigram.

Quantity and Price of a Commodity (1965—1970)

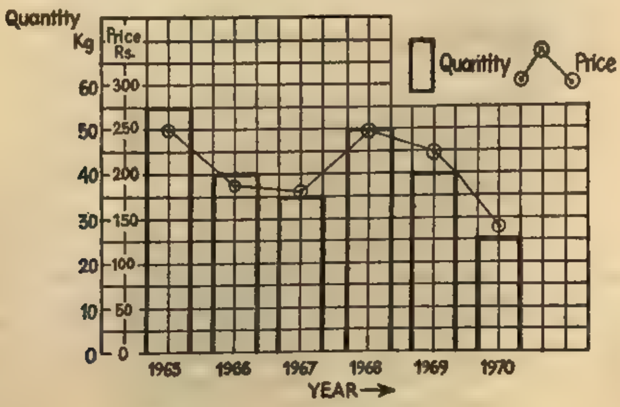


Fig. 6

From the above graph, the variations of quantity and the price, year to year, can be studied. The two variables move in the same direction.

(2) **Zone Graph.**

Sometimes it becomes necessary to represent the maximum and minimum values of a given set of variables by means of a graph. In such case, we are to put two marks—one for the maximum and the other for the minimum value, on the appropriate data. The space between these points are made prominent by thickening the lines.

Example.

Average prices of gold in a certain city (per tola) in terms of Rupees, are shown below—

Year	1936	'37	'38	'39	'40
Maximum Price (Rs.)	36·75	35·50	35·20	37·70	38·10
Minimum Price (Rs.)	31·25	33·94	34·25	34·75	35·12

Maximum and Minimum Prices of Gold

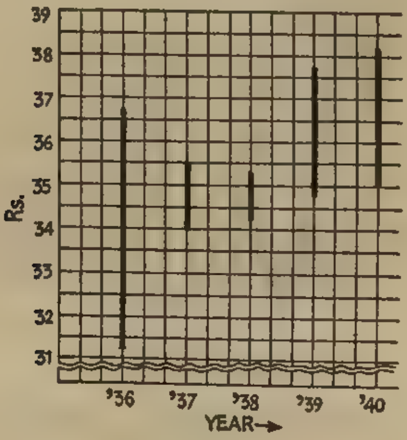


Fig. 7

(3) **Band Curve.**

It is a type of linear graph, used to represent the total of component parts of some data spread over a period of time. The components may be plotted one above the other, using different shades in the space formed. This type of graph is useful for studying total cost divided in various component parts.

Example.

Represent the following data graphically—*Advances granted by Primary Agricultural Credit Societies* (*in crores of Rs.*)

Year (1)	Bombay (2)	Madras (3)	All-India (4)
1946-47	1·70	3·47	9·03
-48	2·22	4·40	10·45
-49	3·29	4·96	14·40
-50	5·29	6·44	17·99
-51	6·90	7·65	22·90
-52	8·12	7·33	24·21

(Source : All India Rural Credit Survey, 1854)

Calculation for Plotting the Data

Year	Col. (2)	Col. (2) + Col. (3)	Col. (2) + Col. (3) + Col. (4)
1946-47	1·70	5·17	14·20
-48	2·22	8·62	19·07
-49	3·29	8·25	22·65
-50	5·29	11·73	29·72
-51	6·90	14·55	37·45
-52	8·12	15·45	39·66

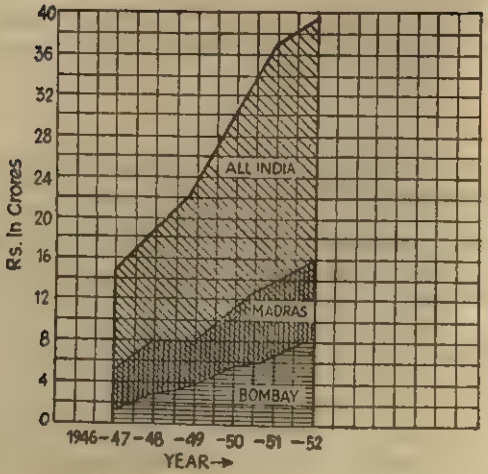


Fig. 8

Ratio Scale.

Logarithmic Charts or Ratio Charts. In the graph drawn before, we have the natural scale, when same number of units in the graph, being represented by same distance. For example, if the pair of numbers 20, 30 and 1000, 1010 are plotted in natural scale (or ordinary) graph, then the vertical distance between the two pairs of points will be the same since each pair differs only by 10.

Again when the variable changes from 20 to 30, there is an increase of 10 *i.e.*, 50% while for the change from 1000 to 1010, the increment is 1%. Now we see that the relative changes are different, though the actual change is the same. Natural scale graph shows only the actual change and not the relative change, while logarithmic graph shows only the relative change and not the actual change.

In order to compare such relative changes over a period of time, a special form of graph known as Logarithmic Graph (or Chart) or Ratio Chart is used. In this graph the vertical axis (*i.e.*, Y-axis) represents the logarithm of the values of the dependent variable, while the horizontal axis (*i.e.*, X-axis) represents the year or months as in the case of arithmetic scale.

For if y be the dependent variable, whose values are represented in Y-axis, then in ratio chart, the value of log y (and not of y) is tabulated.

Logarithmic Chart is also sometimes called semi-logarithmic chart since the vertical scales are logarithmic and the horizontal scales remain the same absolute (or arithmetic). Equal vertical changes on a logarithmic chart represent equal percentage changes and not equal actual changes.

Characteristics.

1. Ratio chart has no zero (point), since it compares the relative changes. A natural scale has zero, since it compares absolute values. Naturally, zero line is essential in case of natural scale but not for the logarithmic scale.

2. They are particularly useful for representing graphically a very wide range of values *i.e.*, for values ranging from 10,000 to 1,0,0,00,000.

3. Equal vertical distances in a logarithmic graph, represent the same relative change, while in case of natural scale graph, they represent equal absolute changes.

4. Natural scale graph can show the negative values while logarithmic graph cannot show negative values since it has no zero point.

5. Logarithmic scale is specially important in the case of index historigrams. They should be generally drawn on ratio scales, because index numbers are more concerned with proportionate changes than with actual ones.

6. Ratio scale makes extrapolation—finding out a future possible figure, if the data are organic in character. For example, if the population figure of a certain country is plotted on a ratio scale, then the curve obtained may be extended in continuation with its trend beyond the last date to the next date, to obtain a fairly accurate estimate of the next figure.

7. A logarithmic chart can be drawn either on an ordinary graph paper by plotting the logarithm values of the dependent variable or by plotting the actual values on a semi-logarithmic graph paper.

Example.

The following table gives the total units produced at the beginning of the different years. Represent the data graphically and estimate the mid-year values for 1949 and 1953.

Year	1947	'48	'49	'50	'51	'52	'53	'54	'55
Units produced.	20	62	147	300	536	811	1104	1425	1755

[I. C. W. A. Jan. 1965]

Since the units produced differ widely *i.e.* 20 in 1947 to 1755 in 1955, the ratio chart is suitable for representation of the above data.

Calculation.

Year	Units (y)	Log y	Log y (approximate)
1947	20	1·3010	1·30
'48	62	1·7924	1·79
'49	147	2·1673	2·17
'50	300	2·4771	2·48
'51	536	2·7292	2·73
'52	811	2·9090	2·91
'53	1104	3·0430	3·04
'54	1425	3·1538	3·15
'55	1755	3·2442	3·24

Semi Logarithmic Graph (drawn on ordinary paper)

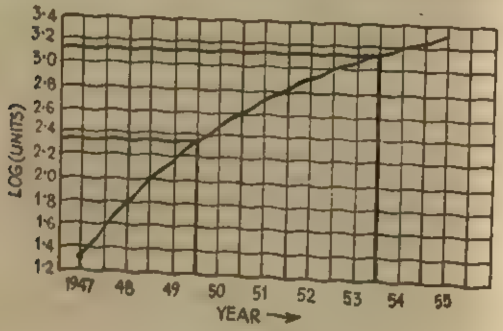


Fig. 9

For estimating mid-year values for 1949 and 1953, two vertical lines are drawn from mid-points of intervals representing the years mentioned, on the horizontal axis until they meet the ratio chart. From the points of intersections, two horizontal lines are drawn so as to meet the vertical axis. Now from the graph, the values of log y are read (app.) as 2·32 and 3·10 whose antilogarithms are 209 and 1259. So the required estimates are 209 and 1259.

GRAPHS OF FREQUENCY DISTRIBUTION

In constructing frequency graphs, the values of the variable are measured on X-axis, while the corresponding frequencies are taken on Y-axis. The following types of graphs can be constructed to represent frequency distribution—

1. **Individual Observation Series Graph :** Let us take any set of data relating to individual items, for example, the followings are the marks obtained by 15 students in a certain class-test of Mathematics.

(Full Marks : 50)

Serial no.	*Marks*	*Serial no.*	*Marks*	*Serial no.*	*Marks*
1	20	6	32	11	39
2	22	7	34	12	40
3	25	8	37	13	42
4	27	9	38	14	42
5	32	10	38	15	46

To represent the data graphically.

Marks obtained by 15 students in Mathematics

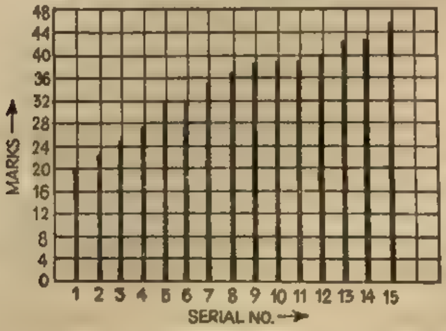


Fig. 10

2. **Discrete Series Graph :** A discrete series is one in which an item cannot assume any value in a class-interval. The value of the item is fixed and definite. This type of series is represented by line or Bar-Frequency Diagrams.

Bars may be vertical or horizontal. The length of the bars is proportional to the values they represent. The base line should be zero, when bar-charts are used for comparison.

Example

Marks	1	2	3	4	5	6	7	8	9	Total
No. of Students	3	5	6	6	7	9	7	6	4	53

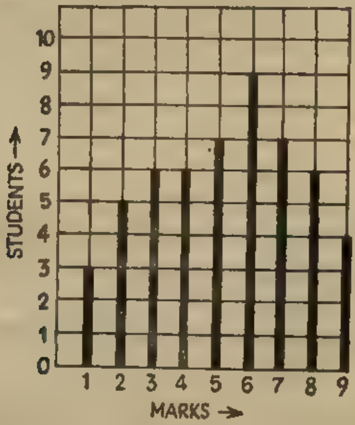


Fig. 11

Note. It may be noted from the above graph that there are no students securing marks between 1 & 2 or 3 & 4 etc. For horizontal bar-graph see Fig. 18.

3. **Continuous Series Graph:** It is one in which an item can assume any value within a particular class-interval.

(i) ***Histogram*** (*when class-intervals are equal*). Let us consider a frequency distribution having a number of class-intervals with their respective frequencies. Now the horizontal axis is marked off to represent the class-intervals and on these markings, rectangles are drawn by taking the lengths of the class-intervals as breadth and corresponding frequencies as heights. Thus a series of rectangles are obtained whose total area represents the total of the class-frequencies. The figure thus obtained is known as *Histogram*.

Example.

Marks	20-30	30-40	40-50	50-60	60-70	70-80	Total
No. of Students	15	55	80	45	25	15	235

The X-axis is marked off suitably to represent the class-intervals of marks (in question). Now, on these class-intervals rectangles are drawn taking the corresponding frequencies (in this case—No. of Students) as height.

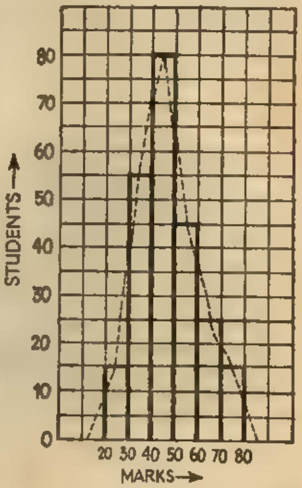


Fig. 12

Note. We can also estimate mode of a frequency distribution by the help of a histogram shown in the chapter of Average.

Histogram (*when class-intervals are unequal*). If the class-intervals are unequal, the frequencies must be adjusted before constructing the histogram. Adjustments are to be made in respect of lowest class-interval. For instance, if one class-interval is twice as wide as the lowest class-interval, then we are to divide the height of the rectangle by two, and if again it is three times more, then we are to divide the height of its rectangle by three and so on.

Example.

Represent the following data by means of a histogram :

Weekly Wages (Rs.)	No. of Workers	Weekly Wages (Rs.)	No. of Workers
10—15	7	30—40	12
15—20	19	40—60	12
20—25	27	60—80	8
25—30	15		

[C. A. 1963]

Since the class-intervals are unequal, frequencies are adjusted as follows—

Weekly Wages (Rs.)	10-15	15-20	20-25	25-30	30-35	35-40	40-45	45-50
No. of Workers	7	19	27	15	6	6	3	3

50-55	55-60	60-65	65-70	70-75	75-80
3	3	2	2	2	2

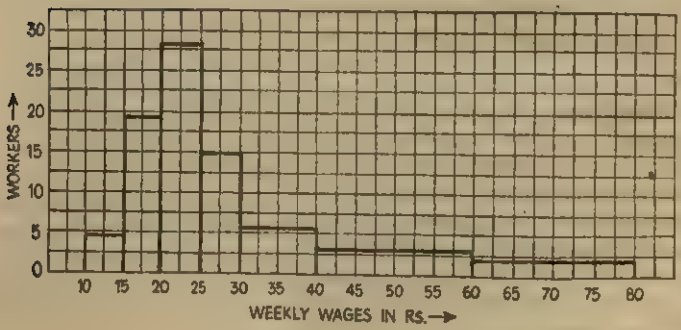


Fig. 13

Histogram (*when only mid-points are given*). When only mid-points (of class-intervals) are given, we are to ascertain the upper and lower limits of the various classes and then to construct the histogram.

Example.

Draw a histogram of the following frequency distributions :

Life of Electric Lamps (in hours) mid-values	1010	1030	1050	1070	1090
Firm	10	130	482	360	18

[I. C. W. A. 1963]

From the mid-values, the class-limits are ascertained as given below :

Life of Electric Lamps	1000-1020	1020-1040	1040-1060	1060-1080	1080-1100
Frequency	10	130	482	360	18

Now the histogram can be drawn easily, similar to Fig. 13.

Histogram (*for discontinuous grouped data*). For discontinuous grouped data, first we are to make class-boundaries (discussed in detail in the chapter of Frequency Distribution) then to draw the histogram by usual method.

Example.

Class-limits	10 – 19	20 – 29	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89
Frequency	5	9	14	20	25	15	8	4

Class-boundaries	9·9 – 19·9	19·9 – 29·9	29·9 – 39·9	39·9 – 49·9	49·9 – 59·9	59·9 – 69·9	69·9 – 79·9	79·9 – 89·9
Frequency	5	9	14	20	25	15	8	4

Now taking the class-boundaries on X-axis, and corresponding frequencies as heights of rectangles, draw the histogram.

(ii) ***Frequency Polygon.**** The line chart obtained by joining successively the middle-points of the tops (uppermost sides) of the rectangles in histogram by straight lines, is known as Frequency Polygon. It is customary to join the extreme two middle-points to the base line at the middle-points of the next class-intervals. The area covered by the frequency polygon is nearly the same as by the histogram.

The dotted line of the Figure (12) represents the Frequency Polygon.

The frequency polygon can also be drawn without the help of a histogram. Points are plotted by taking the middle-points of the

* Polygon literally means *many angles*. In statistics it means a curve representing a frequency distribution.

class-interval as abscissa (x-coordinate) and the corresponding frequency as ordinate (y-coordinate). Then the line chart obtained by joining such points by straight lines is known as Frequency Polygon.

Example.

Class-intervals (marks)	Mid-points	No. of Students
20-30	25	15
30-40	35	55
40-50	45	80
50-60	55	45
60-70	65	25
70-80	75	15
Total	...	235

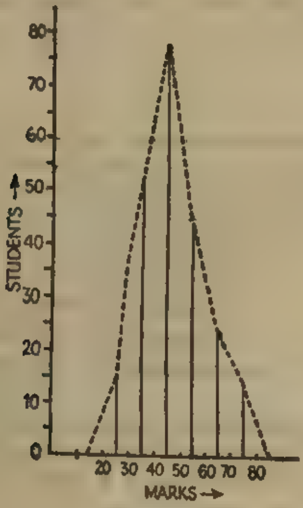


Fig. 14

(iii) ***Frequency Curve.*** The frequency polygon consists of sharp turns, ups and downs. To remove these sharps features of a polygon, it becomes necessary to smooth it. There is no definite rule for smoothing the polygon. Figure (14) shows the smoothed curve (*i.e.* frequency curve).

(iv) ***Ogive or Cumulative Frequency Polygon.*** If the cumulative frequencies are plotted against the class-boundaries and the successive points are joined by straight lines, we get what is known as Ogive (or cumulative frequency polygon). There are two types of Ogive :

(a) *Less than type*—Cumulative frequencies *from below* are plotted against the upper class-boundaries.

(b) *Greater than type*—Cumulative frequencies *from above* are plotted against the corresponding lower boundaries.

The former is known as *less than type*, because the ordinate of any point on the curve (obtained) indicates the frequency of all values less than or equal to the corresponding value of the variable represented by the abscissa of the point. Similarly, the latter one is known as *greater than type.*

Frequency distributions of marks obtained by 170 students

Class-intervals (marks)	Class-boundaries	Frequency	Cumulative Frequency	
			from below (less than 30·5, 40·5 etc.)	from above (greater than 20·5, 30·5 etc.)
21—30	20·5—30·5	15	15	170
31—40	30·5—40·5	25	40	155
41—50	40·5—50·5	40	80	130
51—60	50·5—60·5	60	140	90
61—70	60·5—70·5	20	160	30
71—80	70·5—80·5	10	170	10
Total	—	170	—	—

Note. From the above figure, it is noticed that the ogives cut at a point whose ordinate is 85 *i.e.*, half the total frequency and the corresponding abscissa is 51·33, which is the median of the above frequency distribution (see the sum on median, in the chapter of Average). Even if one ogive is drawn, the median can be determined by locating the abscissa of the point on the curve, whose cumulative frequency is $\frac{N}{2}$. Similarly, the abscissa of the points on the *less*

Ogive of Marks Obtained by 170 Students

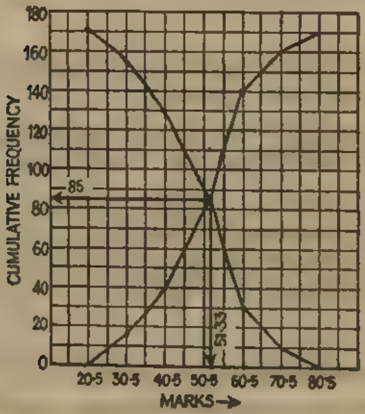


Fig. 15

than type ogive corresponding to the cumulative frequencies $\frac{N}{4}$ and $\frac{3N}{4}$ give the Q_1 (first quartile) and Q_3 (third quartile) respectively. (Q_1, Q_3 will be discussed after median in the chapter of Average).

Example.

The following table gives the average earnings of the mill-workers in a certain city :

Monthly Wages (in Rs.)	Frequency	Monthly Wages (in Rs.)	Frequency
18	21	42	36
21	29	45	45
24	19	48	27
27	39	51	48
30	43	54	21
33	94	57	12
36	73	60	5
39	68		

Draw a histogram and a frequency curve for the data given above. Find the number of mill-workers whose wages lie between Rs. 31 and Rs. 53.

[B. Com. Madras 1962]

We are to make the data in the form of a frequency distribution with class-intervals, as shown below.

Monthly Wages (Rs.)	Frequency	Monthly Wages (Rs.)	Frequency
18-21	21	42-45	36
21-24	29	45-48	45
24-27	19	48-51	27
27-30	39	51-54	48
30-33	43	54-57	21
33-36	94	57-60	12
36-39	73	60-63	5
39-42	68		

Now the histogram can be easily drawn, for reference see. Fig. 12.

For the second part, we are to make cumulative frequèncy distribution as follows—

Monthly Wages (Rs.)	Cum. Frequ. (less than type)	Monthly Wages (Rs.)	Cum. Frequ. (less than type)
less than 21	21	less than 45	422
,, ,, 24	50	,, ,, 48	467
,, ,, 27	69	,, ,, 51	494
,, ,, 30	108	,, ,, 54	542
,, ,, 33	151	,, ,, 57	563
,, ,, 36	245	,, ,, 60	575
,, ,, 39	318	,, ,, 63	580
,, ,, 42	386		

From the Graph [Fig. 16], it is clear, that the number of workers, whose wages lie between Rs. 31 and Rs. 53 is 530-120 *i.e.* 410.

Note. For drawing the curve from *less than* 18, we are to start from X-axis *i.e.* the ogive should start from X-axis itself.

Cumulative Frequency Curve

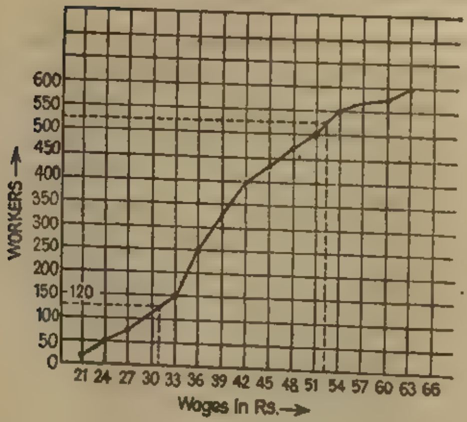


Fig. 16

DIAGRAMMATIC REPRESENTATION

Data may also be represented in the form of a surface figure (*i.e.* in the form of a diagram), other than a curve or graph. For drawing a graph or curve, usually a graph paper is required, whereas in case of a diagram, plain paper may be used.

Types of Diagrams.

The followings are the important types of diagrams, common in use :

(*i*) *One dimensional diagrams i.e.* lines or bars drawn to a common scale.

(*ii*) *Two dimensional diagrams i.e.* rectangles, squares and circles, whose areas are made proportional to the given figures.

(*iii*) *Three dimensional diagrams i.e.* cubes, cylinders or blocks whose volumes are made proportional to the given figures.

(*iv*) *Pictograms i.e.* statistical pictures.

(*v*) *Cartograms i.e.* statistical maps.

Directions for Drawing Diagrams.

It is mentioned before that diagrams give a pictorial representation of quantitative data and show nothing beyond it. Diagrams are not suitable for further analysis of data, which can be done from figures. Before drawing a diagram, one must be sure that the data are capable of diagrammatic representation. All types of data cannot be represented by diagrams. Statistical data should be homogeneous and comparable for such representation. For instance, a set of figures relating to a number of sheets, a number of patients, and a number of schools in relations to our country cannot be represented by means of diagrams. The figures are entirely unrelated. A single figure is also useless for such presentation.

Another point that should be kept in mind is that diagrams are not the substitutes of the real magnitude of the quantity they represent. The size of the diagrams changes with the change in the scale to which it is drawn. The same data drawn in two different scales will yield diagrams of different sizes. The scale of the diagrams should be appropriate and also should suit the size of the paper on which it is drawn. The scale should always be indicated in the figure, as without it, no diagram is complete. A good diagram should be neat, clean and appealing to eye. Each diagram should have a proper heading.

All types of diagrams are not suitable to represent all types of data. It is essential to select that diagram which suits best to represent the data given, otherwise misleading impressions may be created.

One Dimensional Diagram

(1) **Simple bar-diagram** consists of a number of bars of uniform width separated by equal intervening spaces. The length of the bars is proportional to the values they represent. The bars may be placed vertically or horizontally. Bar-diagram is generally used to represent a time-series. The base line should be the zero-line, when bar-diagrams are used for comparison.

Example.

The monthly productions of bi-cycles in a factory are as follows—

January—70, February—60, March—90, April—80, May—100, June—110. —Represent by simple bar-diagram.

Scale : 1 division along Y-axis = 10 units.

Monthly Productions of Bi-cycles

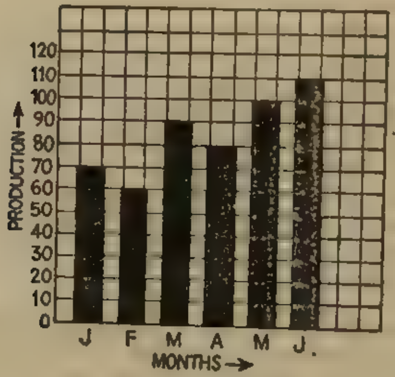


Fig. 17

Example. Construct a horizontal bar-diagram showing expenditure of First Five-year Plan in West Bengal.

	(Crores of Rs.)
On Industries	110'00
On Irrigation	67'50
On Agriculture	90'00
On Transports and Roads	42'50
On Miscellaneous	50'00

Scale : 1 division along X-axis = 10 crores of Rs.

Expenditure in first five-year plan in West Bengal

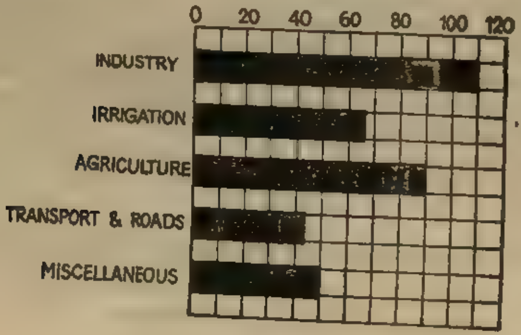


Fig. 18

(2) **Multiple (or Compound) bar-diagrams.** The technique of simple bar-diagrams may be extended to represent two or more sets of inter-related data in one diagram. So multiple bar-diagrams supply information of more than one phenomena.

Example.

Population of Men and Women in districts of North Bengal, according to the Census 1961 are shown below :

Districts	*Men*	*Women*
Darjeeling	3,34,553	2,90,326
Jalpaiguri	7,32,590	6,27,520
Cooch Behar	5,39,798	4,79,953
West Dinajpur	6,96,759	6,33,587
Malda	6,22,092	5,98,399

—Represent the data by multiple bar-diagrams.

Scale : 2 divisions along Y-axis = 1 lakh.

Population in the districts of North Bengal

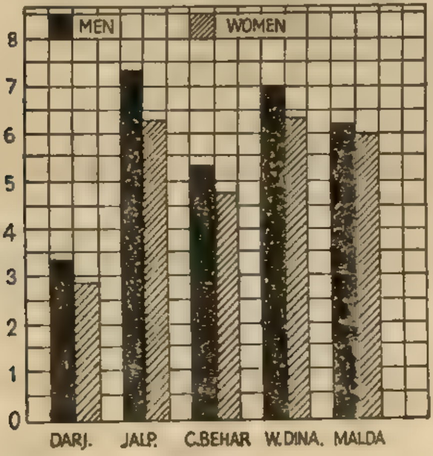


Fig. 19

Example.

Allotment of Money to West Bengal in first three Five-year Plans are as follows—

Five-year Plan	1	2	3
Rs. (in crores)	70	155	340

A triple bar-diagram showing allotment of Money to West Bengal in three Five-year Plans

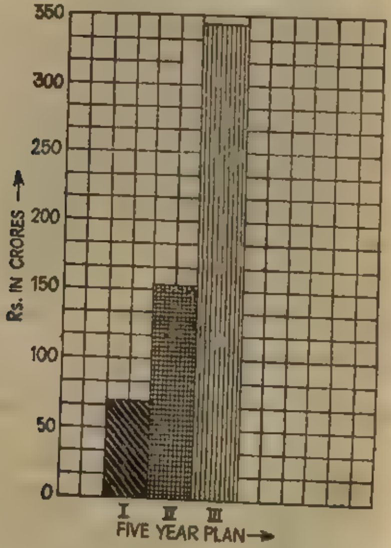


Fig. 20

(3) **Component bar-diagram** (*or Sub-divided bar-diagram*): As the name suggests, these diagrams show the divisions of a whole into its component parts. Component bar-diagrams exhibit in a striking manner the relation between the different parts and also between the parts and the whole.

Example.

The following table shows the total cost (in rupees) and its component parts in two consecutive years.

	1970	*1971*
Direct material	4,000	5,500
Direct labour	5,000	6,000
Direct expenses	1,200	1,500
Overhead	2,500	2,000
	12,700	15,000

Component bar-diagram showing total cost and its components

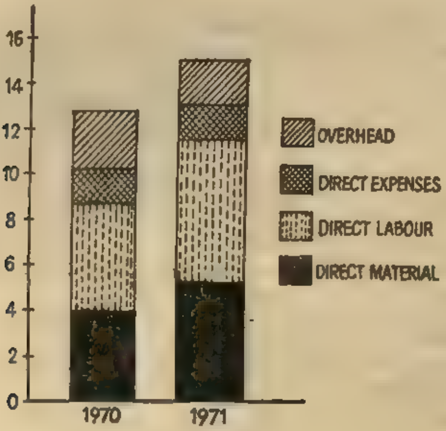


Fig. 21

(4) **Sub-divided bar-diagram on percentage basis.** Comparison of the related data by the above process may be misleading in some cases. A proper and fair comparison may be possible by placing the related data in the same footing Here items constituting the aggregate are expressed as percentages to the aggregate. The length of the bar is equal to 100, and from this, sub-divisions are made according to the percentages they bear to the aggregate, to represent the components. This helps comparison very simple and clear. Use different shades for different components.

Example.

The cost, sale proceeds and profit (or loss) per chair during 1967, '68, '69 are given below :

Particulars	*1967*	*1968*	*1969*
Wages	8	8	11
Other costs	6	5·6	6
Polishing	4	6·4	5
Total cost	18	20	22
Sale proceeds (per chair)	20	20	20
Profit (+) or loss (−) (per chair)	(+)2	nil	(−)2

To represent the above data by sub-divided bar-diagrams on percentage basis.

Before constructing the diagram, we are to convert the quantities into percentages of the sale proceeds as follows :

Particulars	*1967*	*1968*	*1969*
	%	%	%
Wages	40	40	55
Other costs	30	28	30
Polishing	20	32	25
Total cost	90	100	110
Sale proceeds	100	100	100
Profit (+) or loss (−) (per chair)	(+)10	nil	(−)10

Percentage of cost, proceeds, profit (or loss) per chair during 1967, '68, '69

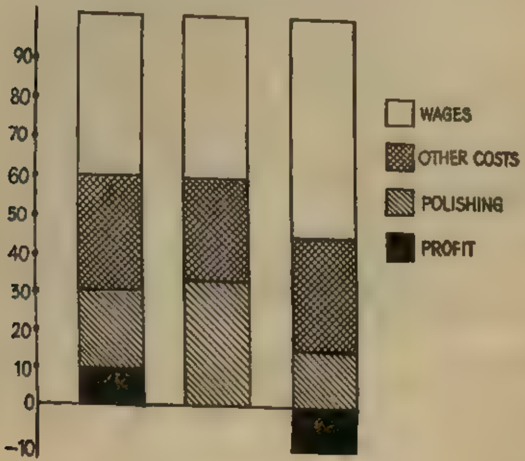


Fig. 22

Two Dimensional Diagram.

(1) **Rectangular diagram.** In one dimensional diagram, as discussed previously, only the length of a bar was taken into account and

not the width. But in two dimensional diagrams, both the length and width are to be taken into consideration. The area of a rectangle is equal to the product of its length and breadth. The area of a rectangle represents the size of the item. The process is similar to that of sub-divided bar-diagrams on percentage basis, except the widths which vary in proportion to the aggregate of each item.

Example.

The student population of the colleges A and B in the different departments are shown below :

	College A	*College B*
Arts	800	400
Science	500	200
Commerce	900	250
Law	300	150

—To represent by the rectangular diagrams.

The aggregates of the students of the two colleges A and B are 2500 and 1000 respectively. So the widths of the rectangles will be proportional to 2500 : 1000 or 5 : 2 (the lengths of the rectangles should be same). Conversion of the figures into percentages of the aggregate are required before constructing the diagrams. The conversions are given below—

	College A		*College B*	
	Students	Percentage	Students	Percentage
Arts	800	32	400	40
Science	500	20	200	20
Commerce	900	36	250	25
Law	300	12	150	15
Total	2500	100	1000	100

Now instead of showing percentages, the actual figures can also be drawn. In such case, the width and length of the rectangles will vary.

Rectangular diagram showing student population of colleges A and B

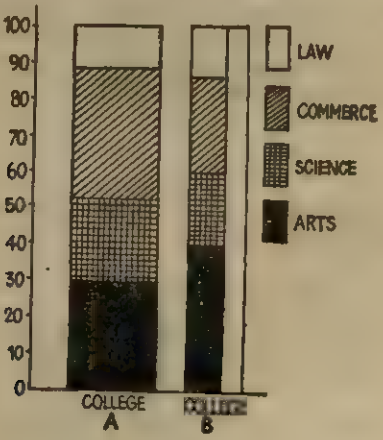


Fig. 23

Example.

Cost of Production, Profits and No. of Units produced by two factories A and B.

Particulars	*Factory A (in Rs.)*	*Factory B (in Rs.)*
Raw materials	100	50
Wages	250	100
Total costs	350	150
Profits	150	90
Total sales	500	240
No. of unit (produced)	100	80

From the above table, sale prices (per unit) of factories A and B are respectively Rs. 5 (500÷100) and Rs. 3 (240÷80). The widths of the two rectangles would be 5 : 3, and the length would be 100 : 80. The rectangles would indicate the total sale-proceeds within which divisions would be done for representating items of costs and profit.

In the first rectangle the items profit, wages and materials would be in the ratio of 150 : 250 : 100 *i.e.*, 3 : 5 : 2 and the vertical scale is divided in such ratios. Similar treatment would be for the second rectangle also.

Cost of Production, Sale-proceeds and Profits of a Commodity in Factories A and B

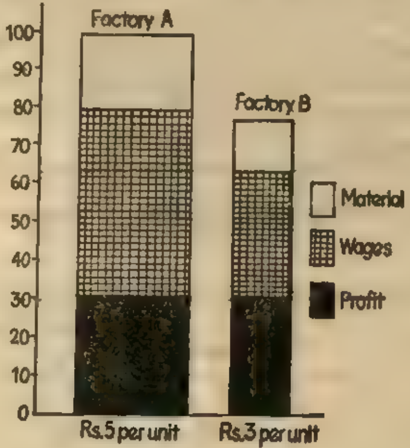


Fig. 24

(2) **Square diagrams.** If it is required to compare quantities in the ratio of 1 : 25, then bar-diagrams become unsuitable, since the height of one bar should be 25 times greater than the other. One bar will be too small, and the other too tall. In such cases, square diagrams give better result.

Now the side of a square varies as the square root of its area. So for representating two figures 100 and 2500 by squares, the sides should be in the ratio of $\sqrt{100} : \sqrt{2500}$ *i.e.* 10 : 50 and not in the ratio of 100 : 2500.

The method is simple. At first the square roots of the given values are taken. The sides of the squares are made in proportion to these square roots. The squares of the sides represent the data. Squares should be placed on a common base and the diagram must be accompanied to the scale of construction.

Example.

Main heading of income of the Central Govt.	1948-49 (in Lakhs of Rs.)
Income tax	13,998
Import duty	7,274
Production duty	5,063
Other taxes	319

It is required to represent the figures by square-diagrams.

Calculation.

Items	Rs. (in Lakhs)	Square roots	Side of square (cms) Col. (3)÷50
1	2	3	4
Income tax	13,998	118·3	2·366
Import duty	7,274	85·29	1·706
Production duty	5,063	71·17	1·423
Other taxes	329	17·86	·357

Income of the Central Govt. (1948-49)

Scale : 1 sq. cm. = Rs. 893·5 lakhs

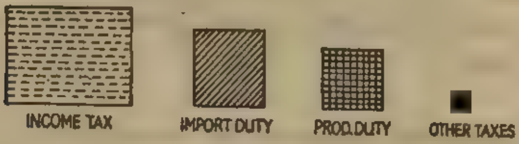


Fig. 25

Calculation of Scale. The area of any square is calculated first. Now the area of a square representing other taxes is ·357 × ·357 sq. cms. This area represents an income of Rs. 319 lakhs. So 1 sq. cm. would represent Rs. 893·5 lakhs.

(3) **Circles.** Circular diagrams are used almost in all those cases, where square diagrams are used. The reason is that area of a circle varies as the square of its radius, so also the area of a square varies with the square of its sides. In construction of circles, like squares the square roots of the various figures are to be calculated and the radii of the circles are kept in the ratio of these square root values.

Example.

The example of previous page (used in squares) is taken, for the construction of circles. Only the Col. 4 is to be read as 'radius of circle', instead of 'side of square'.

Calculation of Scale. The area of the last circle representing the income from 'other taxes' is about ·415 sq. cm. would represent Rs. 319 lakhs, then 1 sq. cm. = Rs. 768·67 lakhs.

Income of the Central Govt. (1948-49)

Scale : 1 sq. cm. = Rs. 768·67 lakhs.

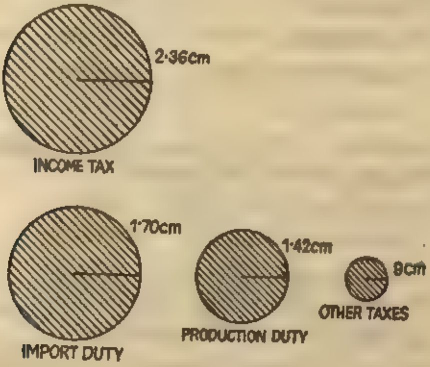


Fig. 26

Circular diagram (or **Pie diagram**). It is a pictorial diagram in the form of circles where whole area represents the aggregate and different sectors of the circle, when divided into several parts, represent the different components.

For drawing a circular diagram, different components are first expressed as percentage of the whole. Now since 100% of the centre of a circle is 360°, 1% corresponds to 3·6 degrees. If p be the percentage of a certain component to the aggregate, then $(p \times 3·6)$ degrees will be the angle, which the corresponding sector subtends at the centre.

Example.

The expenditure during Second Five-year Plan in West Bengal is shown at the next page—

	(Rs. in Crores)
On Industries	127·00
" Irrigation	92·50
" Agriculture	100·00
" Transports & Roads	92·50
" Miscellaneous	68·00
	480·00

—To represent the data by circular diagram.

First we express each item as percentage of the aggregate.

$$\text{Industries} = \frac{127\cdot00}{480\cdot00} \times 100 = 26\cdot4.$$

Irrigation	= 19·3
Agriculture	= 20·8
Transports & Roads	= 19·3
Miscellaneous	= 14·2

Now 1% corresponds to 3·6 degrees. So the angles at the centre of the corresponding sectors are (in degrees)

Industries = 26·4 × 3·6	= 95·0
Irrigation = 19·3 × 3·6	= 69·5
Agriculture = 20·8 × 3·6	= 74·9
Transp. & Rds. = 19·3 × 3·6	= 69·5
Miscellaneous = 14·2 × 3·6	= 51·1

Now with the help of compass and protector (or diagonal scale) the diagram is drawn.

Expenditure during Second Five-year Plan in West Bengal

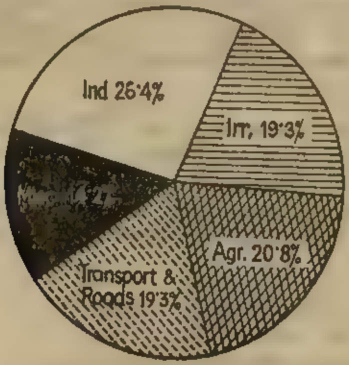


Fig. 27

Note. Additions of all percentages of the items should be equal to 100 and also the addition of all the angles should be equal to 360° (app.).

If two aggregates with their components are to be compared, then two circles are required to be drawn having areas proportionate to the ratio of the two aggregates.

Example. The production cost of two manufacturers A and B.

Particulars	Manufacturer A (Rs. in thousand)	Manufacturer B (Rs. in thousand)
Material	27·7	52·2
Wages	37·7	60·5
Expenses	16·4	32·4
Fact. overhead	18·2	42·9
Total	100·0	188·0

The radii of two circles should be in the ratio of $\sqrt{100}$ and $\sqrt{188}$ *i.e.*, 10 : 13·71 *i.e.*, 10 : 14 (app.) *i.e.*, 5 : 7.

Calculation.

Particulars	Manufac. A		Manufac. B	
	Rs. ('000)	degrees	Rs. ('000)	degrees
Material	27·7	100	52·2	100
Wages	37·7	136	60·5	116
Expenses	16·4	59	32·4	62
Fact. overhead	18·2	65	42·9	82
Total	100	360	188	360

Production cost of two manufacturers A and B

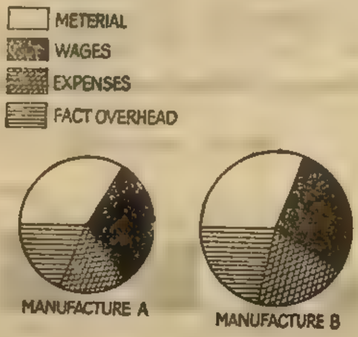


Fig. 28

Three Dimensional Diagrams.

Cubes. In calculating the volume of a cube, three dimensions —length, breadth and depth are to be counted. Hence cube is a three dimensional diagram, and is also known as *volume diagram*, whereas the two dimensional diagrams discussed previously, are known as *surface diagrams.*

Example.

The following table gives the amount deposited in a new branch office of a certain bank for the first four months.

Month	(Rs. in thousand)
1	12
2	70
3	150
4	270

For representing the above figures in cubes. We are to make the cubic roots of the figures, the sides of the cubes should be in proportion to the ratios of cubic roots (if necessary, may be divided by a common factor).

Calculation.

Month	(Rs. in '000)	Cubic roots	Side of cubes (cm.) Col. (3)÷3
(1)	(2)	(3)	(4)
1	12	2·29	·76
2	70	4 12	1·37
3	150	5·31	1·77
4	170	6·47	2·16

Deposits in 4 months

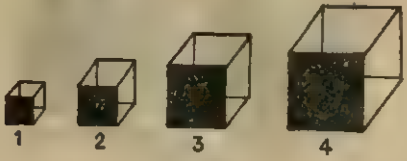


Fig. 29

Note. Cylinders, spheres are also known as three dimensional diagrams. These diagrams are not discussed here, since difficult calculations are required for such constructions.

Pictograms.

The pictures (*i.e.* pictograms) are used very popularly for representing data. The method is quite effective and has the advantage of being easily understood by a common man. Each symbol of picture represents a definite numerical value. If a fraction of the numerical value, represented by a symbol occurs, then the proportionate part of the picture from the left-hand is drawn.

Example.

The table below shows the number of Primary Schools in a certain district for the years 1950, 1960, and 1970.

Years	1950	1960	1970
No. of Primary Schools	20	90	230

Number of Primary Schools in 1950, '60, '70

One figure represents 10 Schools

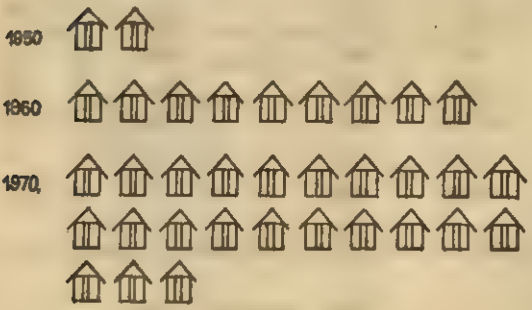


Fig. 30

Cartograms.

Cartograms or statistical maps are used to represent quantitative data on a geographical basis. The quantities on the map are shown through shades, colours, or by pictograms. The maps should be used only where geographical comparisons are of primary importance. The drawing is not shown here, since maps are not common in use.

EXERCISE 3

1. Explain clearly between natural scale and the logarithmic scale used in graphical presentation of data. [I.C.W.A. Jan. 1971]

2. Write short notes on (a) Historigram (b) Ogive (c) Frequency Curve (d) Frequency Polygon.

3. Represent graphically the following data, relating to cheque clearance :

Cheque clearance (Crores of Rs.)

Month Year	Jan.	Feb.	Mar.	Apr.	May	Jun.	July	Aug.	Sep.	Oct.	Nov.	Dec.
1958	832	765	873	792	791	663	834	754	806	799	773	887
1959	894	828	946	946	849	—	—	—	—	—	—	—

[C. U. M. Com. 1961]

4. Represent the following information graphically and also draw a graph on the same sheet to show the balance of trade.

Indian Export and Import in millions of rupees

	Period	Import	Export		Period	Import	Export
1946	April	217	213	1947	Jan.	325	364
	May	218	304		Feb.	320	255
	June	205	254		March	336	307
	July	263	238		April	360	258
	Aug.	227	211		May	409	362
	Sept.	289	200		June	385	354
	Oct.	299	259		July	436	286
	Nov.	313	253				
	Dec.	325	330				

[B. Com. Madras]

5. Represent the following data about a country by a suitable graph :

Production in million tonnes

Year	Rice	Wheat	Pulses	Other cereals
1962	30·4	10	9	16
63	32	11	10	18
64	33	8·5	11·5	20
65	35	12	11	23
66	36·5	10	10	21
67	38	11	9	24

6. The following table shows the foreign trade of Japan. Represent the figures by suitable graph.

Foreign Trade (value : million dollars)

Year	Exports	Imports	Excess of Imports
1940	857	809	48
46	103	306	—202
50	820	974	—154
55	2011	2471	—461
60	4055	4491	—437
65	8452	8169	283
68	12972	12987	—16
69	15990	15024	966

(Source : Ministry of Finance, Japan)

7. What is false base line ? Under what circumstances should it be used ? The following data gives the index number of industrial profits in India. Represent it graphically.

Year	*Index Number* (1929 = 100)	*Year*	*Index Number* (1929 = 100)
1941	187	1946	229
42	222	47	192
43	246	48	260
44	239	49	182
45	234	50	247

8. Explain what is a semi-logarithmic graph. What purpose is served by such graphs and what are its uses ? [I. C. W. A. Jan. 1969]

9. From the following table, draw a ratio chart on a graph paper :

Year	1937	'38	'39	'40	'41	'42	'43	'44	'45
Units Produced	2	4	8	16	32	64	128	256	512

[C. U. M. Com. 1956]

10. Represent graphically from the following data the growth of the population of a state in India to show both relative growth and growth by absolute amount :

Census year	1871	'81	'91	1901	'11	'21	'31	'41
Population (in lakhs)	50·0	52·4	55·6	59·5	65·0	68·7	73·1	77·5

11. The profits of a particular firm and of the whole industry are given in the following table :

Year	1950	'51	'52	'53	'54	'55	'56
Firm (in Rs. 10,000)	1·50	2·00	2·18	2·69	3·50	6·00	14·00
Industry (in Rs. 10,00,000)	3·20	4·00	5·40	6·80	8·00	11·00	19·00

Compare trends of profit by semi-logarithmic graphs and comment on the performance of the firm in relation to that of the industry.

[I. C. W. A. Jan. 1969]

12. The following table shows the values of a variable y corresponding to some given equidistant values of the independent variable x :

x.........	7	8	9	10	11	12
y.........	132	214	330	486	688	942

Draw a semi-logarithmic chart and find by graphical interpolation the value of y, when $x=10·5$

[I. C. W. A. Jan. 1971]

13. (a) What is meant by histogram ? Construct a histogram from the following :—

class limit	90-100	100-110	110-120	120-130	130-140	140-150	150-160
frequency	16	22	45	60	50	24	10

(b) Calculate the number of cases between 112 and 134

(c) Number less than 112

(d) Number greater than 134

[C. A. May 1969]

14. The population of six states in India are given in the following table. Represent the data by bar-diagram.

States	*Population*
Uttar Pradesh	88341144
Bihar	56353369

States	*Population*
Maharashtra	50412235
West Bengal	44312011
Andhra Pradesh	43502708
Tamil Nadu	41199168

[Source : Census 1971]

15. Represent the following data by suitable diagram :

Educated (graduate and post-graduate) unemployed in India

Year	1969	1970	1971	1972 (as on 30. 6. 72)
Number	186,436	232,250	333,421	463,519

[Source : Register Employment Exchange]

16. Describe the advantages of diagrammatic representation of statistical data. Name the different types of diagrams commonly used, and mention the situations where the use of each type of diagram would be appropriate. [I. C. W. A. June 1975]

17. Represent the following data by a suitable diagram showing the difference between proceeds and costs :

Proceeds and Costs of a Firm (in thousands of Rupees)

Year	Total Proceeds	Total Costs
1950	22'0	19'5
51	27'3	21'7
52	28'2	30'0
53	30'3	25'6
54	32'7	26'1
55	33'3	34'2

[I. C. W. A. Jan. 1966]

18. The following table gives the average approximate yield of rice in lbs. per acre in various countries of the world in 1938-39.

Country	India	Siam	U.S.A.	Italy	Egypt	Japan
Yield in lbs. per acre.	728	943	1,469	2,903	2,153	2,276

Indicate this by a suitable diagram which will highlight the relative backwardness of India in this regard. [I. C. W. A. Jan. 1964]

19. Represent the following table by sub-divided bars drawn on a percentage basis :

Cost proceeds and profit or loss per table

Particulars	1951	1956
Cost per table—	Rs.	Rs.
(a) Wages	21	9
(b) Other costs	14	6
(c) Polishing	7	3
Total costs	42	18
Proceeds per table	40	20
Profit (+) or loss (−) per table	(−) 2	(+) 2

[B. Com. Allahabad]

20. Represent the information contained in the following table in a component bar-diagram :

Commodity Pattern of India's Exports (*Percentage*)

	1956-57	1957-58	1958-59
Capital goods	0·29	0·31	0·30
Intermediate goods	45·82	46·87	44·19
Consumer goods	50·50	47·32	48·19
Unclassified	3·39	5·50	7·32
Total	100·00	100·00	100·00

[C. U. B. Com. (Hons.) 1967]

21. The following table shows the number of bushels of wheat and corn produced in a farm during the years 1950 to 1960.

Express the yearly number of bushels of wheat and corn as percentages of total annual production. Graph the percentages by component bar-diagrams.

Year	1950	51	52	53	54	55	56	57	58	59	60
No. of bushels of Wheat	200	185	225	250	240	195	210	225	250	230	235
No. of bushels of Corn	75	90	100	85	80	100	110	105	95	110	100

(Dig. Soc. Welfare, May 1968)

22. The following table shows the monthly expenditure of two families. Represent the data by means of two-dimensions rectangular diagrams.

Item of Expenditure	*Family A* Income Rs. 600 p.m.	*Family B* Income Rs. 1000 p.m.
Food	200	360
Clothing	80	120
House Rent	120	200
Education	100	120
Miscellaneous	50	100

(*Hints* : Savings will be Rs. 50 & Rs. 100 respectively for the two families.)

23. Draw a suitable diagram to represent the following information :

Factory	*Wages* (Rs.)	*Materials* (Rs.)	*Profits* (Rs.)	*Units produced*
A	2,000	3,000	1,000	1,000
B	2,200	2,400	1,000	800

—Show also the cost and profit per unit [B. Com. Allahabad]

24. Represent by square diagrams the following data :

Educated (under graduate) unemployed persons in India

Year	1969	1970	1971	1972 (as on 30. 6. 72)
No. of persons in '000	356	395	529	663

[Source : Employment Exchange]

25. Construct a pie-diagram for the following data :

Principal Exporting Countries of Cotton

(1,000 bales)—1955-56

U. S. A.	*India*	*Egypt*	*Brazil*	*Argentina*
6,367	2,999	1,688	650	202

[C. U. M. Com. 1959]

26. Of the Life Insurance policy dividends paid in the United States 21% were taken in cash, 31% were used to pay premiums, 18% were used to purchase additional paid-up Life Insurance, 30% were left with Life Insurance companies to earn interest. Construct a pie-diagram showing these different uses of policy dividends.

[C. U. M. Com. 1962]

27. The following data represent the share of important producing states in the total area and production of Tobacco in India, during the year 1957-58.

Draw two pie-diagrams to represent the informations :

States	*Percentage*	
	Area	Production
Andhra	39·1	43·3
Bombay	25·5	20·2
Mysore	11·1	10·7
West Bengal	4·5	4·4
U. P.	4·5	3·6
Others	15·3	17·8
Total	100·0	100·0

[Source : Tobacco India, 1957-58]

28. Represent the following figures by cubes :

Number of Students during 1971-72, in India

	Primary Schools	All Schools	Universities
(in Lakhs)	686	840	24

[Source : Census 1971]

29.

Year	1863	1864	1865	1866	1867
No. of Fact. in European Russia	11800	12000	13700	6900	7100

—Present the above data in a graph. [I. C. W. A. June 1979]

30. The table below shows the exports of woven-piece goods in Million Square Yards during some months in a year :

	Cotton	*Wool*
April	96	15
May	78	10
June	72	9
July	65	10
August	77	10

Make a graphical comparison of the compound bar-chart of the volume of exports given in the above table.

[I. C. W. A. June 1979]

31. Draw histogram and frequency polygon to present the following data :

Income (*Rs.*)	*No. of Individuals*
100 – 149	21
150 – 199	32
200 – 249	52
250 – 299	105
300 – 349	62
350 – 399	43
400 – 449	18
450 – 499	9
	342

[I. C. W. A. June 1978]

32. Discuss the types of data which are usually represented by pie-diagrams. State how they are drawn.

Represent the following data by a bar-diagram :

Production of Sugar in a certain year

	in quintals (000, 000)
Cuba	32
Australia	30
India	20
Japan	5
Java	1
Egypt	1
	89

[I. C. W. A. June 1978]

33. Draw a pie-diagram to represent the following population in a town :

Males	*Females*	*Girls*	*Boys*	*Total*
2,000	1,800	4,200	2,000	10,000

[I. C. W. A. Dec. 1977]

34. Draw the graph of the following :

Year	1920	'21	'22	'23	'24	'25	'26	'27
yield (in million tons)	12·8	13·9	12·8	13·9	13·4	6·5	2·9	14·8

[I. C. W. A. Dec. 1977]

35. Explain the use of various diagrams in presenting statistical data. [I. C. W. A. June 1976]

36. The following data show the estimated savings of the household sector in India during 1662-63, as revealed by the C.S.O.

Form of Savings	*Amount (Rs. crores)*
Currency	175
Provident Fund	145
Physical	158
Others	440

Present the information in a suitable diagram so as to enable comparison among the various components and also in relation to the total. [C. U. B. Com (Hons.) 1980]

37. A ship has four compartments labelled 1, 2, 3 and 4. The space limits of 1, 2, 3 and 4 are respectively 180,000 cubic feet, 160,000 cubic feet, 140,000 cubic feet and 120,000 cubic feet. Present the data about the different space limits in a table and draw a Pie diagram to represent the above data. [I. C. W. A. June 1980]

5

INTERPOLATION

Introduction

The term interpolation means framing the most appropriate estimate of a missing quantity, under certain reasonable estimate.

In the Chapter of Average (discussed later on), the median and mode were interpolated in the median and modal classes respectively. This, of course, was done by proceeding with certain assumptions. Again let us take the following inter-related values of two variables x and y.

x	1	2	3	4	5
y	10·1	11·5	12·0	13·6	14·7

Now for $x=3$, $y=12{\cdot}0$ and for $x=4$, $y=13{\cdot}6$, but we do not know the value of y variable for x is 3·6. The technique of estimating the value of y for x is 3·6, would be called interpolation. Again the technique of estimating a past figure is termed as interpolation, while that of estimating a probable figure for the future is called *extrapolations.*

Assumptions.

We cannot supply the missing figure just arbitrarily, but we are to make the most appropriate estimate. The making of this appropriate requires certain assumptions, which are as follows—

1. There are no sudden jumps from one period to another *i.e.*, distribution should be normal. If, however, there are violent disturbances, the estimation of interpolation will be impossible.

2. The rate of change of figures from one period to another is uniform.

Finite Difference

In problem of interpolation, the independent variable is known as *argument*, while the dependent variable is usually called a *function* (or entry) of the former. Let $x_0, x_1, x_2, \ldots\ldots, x_n$ are successive values of argument having a constant increment, and $y_0, y_1, y_2, \ldots\ldots, y_n$ are the corresponding functions, then $(y_1 - y_0), (y_2 - y_1), \ldots\ldots, (y_n - y_{n-1})$ are called *finite differences of first order* or (simply *first differences*). These differences are denoted respectively by $\Delta y_0, \Delta y_1, \ldots\ldots, \Delta y_{n-1}$ *i.e.*, $\Delta y_0 = y_1 - y_0$, $\Delta y_1 = y_2 - y_1$, $\ldots\ldots$, $\Delta y_{n-1} = y_n - y_{n-1}$ proceeding similarly with the first differences, we have $(\Delta y_1 - \Delta y_0)$, $(\Delta y_2 - \Delta y_1), \ldots\ldots, (\Delta y_{n-1} - \Delta y_{n-2})$ which are known as *finite differences of second order* (*or second differences*), denoted respectively by

$$\Delta^2 y_0, \Delta^2 y_1, \ldots, \Delta^2 y_{n-2}, \textit{ i.e.,}$$
$$\Delta^2 y_0 = \Delta y_1 - \Delta y_0, \Delta^2 y_1 = \Delta y_2 - \Delta y_1, \ldots\ldots$$
$$\Delta^2 y_{n-2} = \Delta y_{n-1} - \Delta y_{n-2}$$

In the same way, third differences $\Delta^3 y$, fourth differences $\Delta^4 y$ etc., may be calculated.

A numerical *example* will make it clear.

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
4	7				
		3			
			1		
6	10			0	
		4			1
8	14		1		
		5		1	
10	19		2		
		7			
12	26				

In the above table, 7 is the leading term and 3, 1, 0, 1 are the leading differences.

The Symbolic Operator E

E stands for a symbolic operator. If $y = f(x)$ is a function of x, the difference between two successive values of x being h, then $Ef(x)$ means the value of the function corresponding to the next higher value of x, *i.e.*, $Ef(x) = f(x+h)$.

The operation E may be repeated and we write,

$$E^2\ f(x)=E(E\,f(x))=E\ (f(x+h))=f(x+2h)$$

$$E^3\ f(x)=E[\,E\ (E\,f(x))]=E\ [f(x+2h)]=f(x+3h)$$

Similarly, $E^n f(x)=f(x+nh)$

Now $\quad \Delta f(x)=f(x+h)-f(x)=E\,f(x)-f(x)=(E-1)f(x)$

or, $\quad \Delta=E-1$ or, $E=1+\Delta$, which means the operation by E is equivalent to the operation by $1+\Delta$.

We know, $\Delta^2 y_0=\Delta y_1-\Delta y_0=(y_2-y_1)-(y_1-y_0)=y_2-2y_1+y_0$

Similarly $\Delta^3 y_0=y_3-3y_2+3y_1-y_0$

$\Delta^4 y_0=y_4-4y_3+6y_2-4y_1+y_0$ and so on.

Now we can write,

$$\Delta y_0=Ey_0-y_0$$
$$\Delta^2 y_0=E^2 y_0-2Ey_0+y_0$$
$$\Delta^3 y_0=E^3 y_0-3E^2 y_0+3Ey_0-y_0$$
$$\Delta^4 y_0=E^4 y_0-4E^3 y_0+6E^2 y_0-4Ey_0+y_0$$

Again, taking the operators only (*i.e.*, removing y_0),

$$\Delta=E-1,\ \Delta^2=E^2-2E+1=(E-1)^2$$
$$\Delta^3=E^3-3E^2+3E-1=(E-1)^3 \text{ and so on.}$$

In general, $\Delta^r=(E-1)^r$

Example.

Express $\Delta^4 y_0$ in terms of $y_0, y_1, y_2, \ldots\ldots$

$$\Delta^4 y_0=(E-1)^4 y_0=(E^4-4E^3+6E^2-4E+1)y_0$$
$$=E^4 y_0-4E^3 y_0+6E^2 y_0-4Ey_0+y_0$$
$$=y_4-4y_3+6y_2-4y_1+y_0.$$

Differences of Polynomial Function.

If y is a polynomial of nth degree, then the consequence differences of higher degree are all zero.

Example.

Find out with the help of E the value of the production for the year 1964 from the table at the next page :

Year	*Production* ('000 tons)
1961	15
62	18
63	20
64	
65	23

We may write the table as follows—

1961	y_0	15
62	y_1	18
63	y_2	20
64	y_3	
65	y_4	23

Since here only 4 values are known, it will be a polynomial of 3rd degree, and hence the consequence differences of 4th degree will be zero *i.e.*, $\Delta^4 y_0 = (E-1)^4 y_0 = 0$,

or, $E^4 y_0 - 4E^3 y_0 + 6E^2 y_0 - 4E y_0 + y_0 = 0$

or, $y_4 - 4y_3 + 6y_2 - 4y_1 + y_0 = 0$

or, $23 - 4y_3 + 6.20 - 4.18 + 15 = 0$

or, $y_3 = 21{\cdot}5$.

$\therefore$ The required estimated production in 1964 will be 21·5 (in '000 tons).

Methods of Interpolation.

There are two types of methods of interpolations. They are—

1. Graphic method.
2. Algebraic method.

GRAPHIC METHOD

This is the simplest of all the methods of interpolation. The statistical data are to be plotted on a graph paper. After this a continuous smoothed curve can be obtained by joining the plotted points. On the X-axis we take the period and on Y-axis the corresponding variable. For the period for which the variable is to be calculated a perpendicular is drawn from the same period (in X-axis) meeting the smoothed curve. From the point where it meets the curve, another perpendicular is drawn on the Y-axis. The point on Y-axis is read off, which is the required value. The idea will be clear from the example.

Example. From the following data determine the population (of a certain city) in 1946 and find also the increase of populations between 1946 and 1936.

Year	*Population*
	(lakhs)
1931	34
41	39
51	45
61	49
71	52

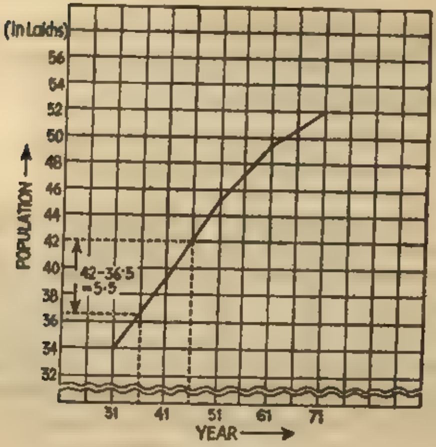


Fig. 31

The graph is drawn by usual process. Now from the graph it is clear that the population of 1946 was 42 lakhs, that for 1936 was 36·5 lakhs and the increase of population was 5·5 lakhs (= 42 − 36·5).

Note. Although it is a simple method, but not an accurate method. Narrower scale is to be taken on the graph for longer volume of figures and consequently the greater will be the error of approximation.

ALGEBRAIC METHOD

Under this method, there are several formulae some of which are given below :

1. Newton's Formula
2. Lagrange's Formula
3. Method of Binomial Expansion

(1) Newton's Formula

Newton's Forward Formula. This formula is suitable when the figure to be interpolated is in the beginning of the table and the values of the argument are equidistant. In the formula, we take only the leading differences into account, and these differences are always in the beginning. *Newton's Forward Formula* is expressed as follows—

$$y_x = y_0 + x\Delta_0^1 + \frac{x(x-1)}{1\times2}\Delta_0^2 + \frac{x(x-1)(x-2)}{1\times2\times3}\Delta_0^3 + \frac{x(x-1)(x-2)(x-3)}{1\times2\times3\times4}\Delta_0^4 + \cdots\cdots$$

where, y_x is the figure to be interpolated, y_0 is the value of origin, Δ's are the differences between adjoining values.

x is calculated as follows—

$$\frac{\text{figure of interpolation} - \text{figure of origin}}{\text{distance between adjoining figures}}$$

Example. The followings are the annual premiums for a policy of Rs. 1,000. Calculate the premium at the age of 32.

Age (in yrs.)	20	25	30	35	40
Premium (Rs.)	24	27	31	36	42·5

Age (yrs.) x		Premium (Rs.)		Differences							
				Δ^1		Δ^2		Δ^3		Δ^4	
20	x_0	24	y_0								
				3	Δ_0^1						
25	x_1	27	y_1			1	Δ_0^2				
				4	Δ_1^1			0	Δ_0^3		
30	x_2	31	y_2			1	Δ_1^2			·5	Δ_0^4
				5	Δ_2^1			·5	Δ_1^3		
35	x_3	36	y_3			1·5	Δ_2^2				
				6·5	Δ_3^1						
40	x_4	42·5	y_4								

$$x = \frac{\text{year of interpolation} - \text{year of origin}}{\text{difference between adjoining values of } x}$$

$$= \frac{32-20}{5} = \frac{12}{5} = 2{\cdot}4.$$

Now putting the above respective values in the formula of Newton we find,

$$y_{32}=24+2{\cdot}4\times3+\frac{2{\cdot}4(2{\cdot}4-1)}{1\times2}\times1+\frac{(2{\cdot}4)(2{\cdot}4-1)(2{\cdot}4-2)}{1\times2\times3}\times0$$
$$+\frac{(2{\cdot}4)(2{\cdot}4-1)(2{\cdot}4-2)(2{\cdot}4-3)}{1\times2\times3\times4}\times({\cdot}5)$$

$$=24+7{\cdot}2+1{\cdot}68+0-{\cdot}00168=32{\cdot}87832=32{\cdot}88 \text{ (app.)}$$

$\therefore$ the reqd. premium = Rs. 32·88

Example.

The following table depicts the number of persons earning certain grades of wages. Estimate the number of persons earning between Rs. 60 and Rs. 70 per mensem :

Wages per mensem (Rs.)	*Number of persons earning* (thousands)
below 40	250
40—60	120
60—80	100
80—100	70
100—120	50

[M. A. (Agra) 1951]

At first we are to estimate the number of persons earning below Rs. 70. From this estimated number we are to deduct the number earning below Rs. 60, for finding required estimated number.

Wages (Rs.)		No. of Persons (cum. fr.)	Differences			
x			Δ^1	Δ^2	Δ^3	Δ^4
below 40	x_0	250 y_0				
			$120\Delta_0^1$			
,, 60	x_1	370 y_1		$-20\Delta_0^2$		
			$100\Delta_1^1$		$-10\Delta_0^3$	
,, 80	x_2	470 y_2		$-30\Delta_1^2$		$20\Delta_0^4$
			$70\Delta_2^1$		$10\Delta_1^3$	
,, 100	x_3	540 y_3		$-20\Delta_2^2$		
			$50\Delta_3^1$			
,, 120	x_4	590 y_4				

$$x=\frac{70-40}{20}=\frac{30}{20}=1{\cdot}5$$

Putting the above respective values in the Newton's formula we get,

$$y_{70}=250+1{\cdot}5\times120+\frac{1{\cdot}5(1{\cdot}5-1)}{1\times2}\times(-20)+\frac{1{\cdot}5(1{\cdot}5-1)(1{\cdot}5-2)}{1\times2\times3}\times(-10)$$
$$+\frac{1{\cdot}5(1{\cdot}5-1)(1{\cdot}5-2)(1{\cdot}5-3)}{1\times2\times3\times4}\times20$$

$$=250+180-7{\cdot}5+{\cdot}625+{\cdot}46875=423{\cdot}59375$$

∴ the reqd. number of earners between Rs. 60 and Rs. 70,

$=423{\cdot}59375-370=53{\cdot}59375$ thousands

$=53594.$

Newton's Backward Formula. This formula is suitable when the figure to be interpolated is lying near the end of the tabulated values, and the values of the argument are equidistant. *Newton's Backward Formula* is expressed as follows—

$$y_x=y_n+x\,\Delta\,y_{n-1}+\frac{x\,(x+1)}{\underline{|2}}\,\Delta^2y_{n-2}+\frac{x\,(x+1)\,(x+2)}{\underline{|3}}\,\Delta^3y_{n-3}$$
$$+\cdots\cdots\cdots+\frac{x\,(x+1)\,(x+2)\ldots(x+n-1)}{\underline{|n}}\,\Delta^ny_0$$

where, y_x is the figure to be interpolated,

y_0 is the value of the origin,

Δ's are the differences between adjoining values.

x is calculated as follows—

$$\frac{\text{figure of interpolation}-\text{figure of last entry}}{\text{distance between adjoining figures}}$$

This is called *Backward Formula* since starting from y_n it uses values of y backward and none forward. The formula uses differences at the bottom of the different columns.

Example.

Find out the value of Y corresponding to X = 18 from the following table :

X	5	10	15	20
Y	7	13	17	18

Let us construct the differences table :

X	Y	Differences		
		Δ	Δ^2	Δ^3
5	7			
		6		
			-2	
10	13			-1
		4		
			-3	
15	17			
		1		
20	18			

Here, $x = \dfrac{18-20}{5} = \dfrac{-2}{5} = -{\cdot}4$

Now, putting the above values in the *backward formula* we find,

$$y_{18} = 18 + (-{\cdot}4) \times 1 + \frac{(-{\cdot}4)(-{\cdot}4+1)}{1.2}(-3) + \frac{(-{\cdot}4)(-{\cdot}4+1)(-{\cdot}4+2)}{1.2.3}(-1)$$

$$= 18 - {\cdot}4 + {\cdot}36 + {\cdot}064 = 18{\cdot}024$$

(2) Lagrange's Formula.

This formula is applicable for the series of unequal intervals. The formula is expressed as follows—

$$y_x = y_0 \frac{(x-x_1)(x-x_2)\ldots(x-x_n)}{(x_0-x_1)(x_0-x_2)\ldots(x_0-x_n)} + y_1 \frac{(x-x_0)(x-x_2)\ldots(x-x_n)}{(x_1-x_0)(x_1-x_2)\ldots(x_1-x_n)} + \cdots\cdots + y_n \frac{(x-x_0)(x-x_1)\ldots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)\ldots(x_n-x_{n-1})}$$

where, y_x is the quantity to be interpolated,

x is the quantity for which the value of y is to be found.

$x_0, x_1, x_2, x_3, \ldots$ are the variables of x-series,

$y_0, y_1, y_2, y_3, \ldots$ are the variables of y-series.

Example.

Determine the percentage of criminals under 35 years of age.

Age	*Percentage of criminals*
Under 25 yrs.	52·0
„ 30 „	67·3
„ 40 „	84·1
„ 50 „	94·4

[B.Com. Nagpur 1963]

Age		*% of criminals*	
Under 25 yrs.	x_0	52·0	y_0
„ 30 „	x_1	67·3	y_1
„ 40 „	x_2	84·1	y_2
„ 50 „	x_3	94·4	y_3

From the Lagrange's formula,

$$y_x = y_0 \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} + y_1 \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)}$$
$$+ y_2 \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} + y_3 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}$$

Putting the respective values in the formula, we find,

$$y_{35} = 52 \frac{(35-30)(35-40)(35-50)}{(25-30)(25-40)(25-50)} + 67{\cdot}3 \frac{(35-25)(35-40)(35-50)}{(30-25)(30-40)(30-50)}$$
$$+ 84{\cdot}1 \frac{(35-25)(35-30)(35-50)}{(40-25)(40-30)(40-50)} + 94{\cdot}4 \frac{(35-25)(35-30)(35-40)}{(50-25)(50-30)(50-40)}$$
$$= 52 \frac{5.(-5)(-15)}{(-5)(-15)(-25)} + 67{\cdot}3 \frac{10.(-5)(-15)}{5.(-10)(-20)} + 84{\cdot}1 \frac{10.5.(-15)}{15.10.(-10)}$$
$$+ 94{\cdot}4 \frac{10.5.(-5)}{25.20.10}$$
$$= -10{\cdot}4 + 50{\cdot}475 + 42{\cdot}05 - 4{\cdot}72 = 77{\cdot}405\%$$

$\therefore$ the reqd. No. of criminals under 35 yrs. = 77·41% (app.)

Example.

Given, log 654 = 2·8156, log 658 = 2·8182, log 659 = 2·8189, log 661 = 2·8202, —find log 656 (all are in common logarithms).

[I.A.S. 1956]

x		y	
log 654	x_0	2·8156	y_0
log 658	x_1	2·8182	y_1
log 659	x_2	2·8189	y_2
log 661	x_3	2·8202	y_3

Using Lagrange's Formula, and substituting the values, we get,

$$y_{656}=2{\cdot}8156\,\frac{(656-658)(656-659)(656-661)}{(654-658)(654-659)(654-661)}$$

$$+2{\cdot}8182\,\frac{(656-654)(656-659)(656-661)}{(658-654)(658-659)(658-661)}$$

$$+2{\cdot}8189\,\frac{(656-654)(656-658)(656-661)}{(659-654)(659-658)(659-661)}$$

$$+2{\cdot}8202\,\frac{(656-654)(656-658)(656-659)}{(661-654)(661-658)(661-659)}$$

$$=2{\cdot}8156\,\frac{(-2)(-3)(-5)}{(-4)(-5)(-7)}+2{\cdot}8182\,\frac{(2)(-3)(-5)}{(4)(-1)(-3)}$$

$$+2{\cdot}8189\,\frac{(2)(-2)(-5)}{(5)(1)(-2)}+2{\cdot}8202\,\frac{(2)(-2)(-3)}{(7)(3)(2)}$$

$$={\cdot}6033+7{\cdot}0455-5{\cdot}6378+{\cdot}8059$$

$$=2{\cdot}8169$$

∴ the required estimated value of log 656 = 2·8169.

Example.

The following table gives the normal weight of a baby during the first six months of life :

Age in months	0	2	3	5	6
Weight in lb.	5	7	8	10	12

Estimate the weight of a baby at the age of 4 months.

[I.C.W.A. Jan. 1970]

Age in months		*Weight in lb.*	
0	x_0	5	y_0
2	x_1	7	y_1
3	x_2	8	y_2
5	x_3	10	y_3
6	x_4	12	y_4

$$y_4 = 5\,\frac{(4-2)(4-3)(4-5)(4-6)}{(0-2)(0-3)(0-5)(0-6)} + 7\,\frac{(4-0)(4-3)(4-5)(4-6)}{(2-0)(2-3)(2-5)(2-6)}$$
$$+8\,\frac{(4-0)(4-2)(4-5)(4-6)}{(3-0)(3-2)(3-5)(3-6)} + 10\,\frac{(4-0)(4-2)(4-3)(4-6)}{(5-0)(5-2)(5-3)(5-6)}$$
$$+12\,\frac{(4-0)(4-2)(4-3)(4-5)}{(6-0)(6-2)(6-3)(6-5)}$$
$$= \frac{(2)(1)(-1)(-2)}{(-2)(-3)(-5)(-6)} + 7\,\frac{(4)(1)(-1)(-2)}{(2)(-1)(-3)(-4)} + 8\,\frac{(4)(2)(-1)(-2)}{(3)(1)(-2)(-3)}$$
$$+10\,\frac{(4)(2)(1)(-2)}{(5)(3)(2)(-1)} + 12\,\frac{(4)(2)(1)(-1)}{(6)(4)(3)(1)}$$
$$= \frac{1}{9} - \frac{7}{3} + \frac{64}{9} + \frac{16}{3} - \frac{4}{3} = \frac{1-21+64+48-12}{9} = \frac{80}{9} = 8\,\frac{8}{9}$$

$\therefore$ the required estimated weight of a baby is $8\,\frac{8}{9}$ lb.

(3) Method of Binomial Expansion.

This method requires some calculations, subject to the satisfaction of the following two points :

(1) The variable x should increase uniformly, say, 2, 4, 6, 8, ... etc. The method is not applicable for otherwise.

(2) The value of x for which y is to be interpolated should be one of the same class-limits of x-series. For example,

x	2	4	6	8	10
y	12	16	?	22	28

We can find the value of y for $x=6$, but not for $x=7$ or 9. Again we can find the value of y for $x=12$ but not for $x=11$ or 13.

Expansion of Binomial, and equating it to zero, we find,

$$(y-1)^n = y^n - ny^{n-1} + \frac{n(n-1)}{1\times 2}\,y^{n-2} - \frac{n(n-1)(n-2)}{1\times 2\times 3}\,y^{n-3} + \cdots = 0,$$

where n is the number of known values of y.

for $n=3$ $\quad y_3 - 3y_2 + 3y_1 - y_0 = 0$

$n=4$ $\quad y_4 - 4y_3 + 6y_2 - 4y_1 + y_0 = 0$

$n=5$ $\quad y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0$

$n=6$ $\quad y_6 - 6y_5 + 15y_4 - 20y_3 - 15y_2 - 6y_1 + y_0 = 0.$

Example.

The following table gives the amount of cement in thousands of tonnes manufactured in the year x. Find the missing term.

x	1956	'58	'60	'62	'64	'66
Cement manf. thousands tonnes	39	85	?	151	264	388

[I.C.W.A. July 1966]

Here, the known values are 5 the fifth leading difference will be 0. Symbolically, it is expressed as

$\Delta^5 = 0$ of which the binomial expansion is expressed as

$$y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0$$

Now,

x	y	
1956	39	y_0
'58	85	y_1
'60	?	y_2
'62	151	y_3
'64	264	y_4
'66	388	y_5

Substituting the values, we find,

$$388 - 5 \times 264 + 10 \times 151 - 10y_2 + 5 \times 85 - 39 = 0$$

or, $388 - 1320 + 1510 - 10y_2 + 425 - 39 = 0$ or, $y_2 = 96{\cdot}4$

$\therefore$ the probable amount is 96·4 thousand tonnes.

Example.

The age of mothers and the average number of children born per mother are given in a table below. Interpolate the average number of children born per mother aged (30—34).

Age of mother (in yrs.)	*Average no. of children born*
15—19	0·7
20—24	2·1
25—29	3·5
30—34	?
35—39	5·7
40—44	5·8

[M.Com. Agra 1968]

Here, $y_0=0\text{·}7$, $y_1=2\text{·}1$, $y_2=3\text{·}5$, $y_3=?$, $y_4=5\text{·}7$, $y_5=5\text{·}8$.

Since the known figures are five, the fifth leading differences will be zero.

Now, $\Delta_0^5=0$

or, $y_5-5y_4+10y_3-10y_2+5y_1-y_0=0$

or, $5\text{·}8-5\times5\text{·}7+10y_3-10\times3\text{·}5+5\times2\text{·}1-0\text{·}7=0$

(substituting values)

or, $5\text{·}8-28\text{·}5+10y_3-35+10\text{·}5-\text{·}7=0$

or, $-47\text{·}9+10y_3=0$ or, $10y_3=47\text{·}9$ or, $y_3=4\text{·}79$

∴ the probable average no. of children = 4·79 or 5 (app.).

EXERCISE 4

1. If l_x represents the numbers living at age x in a life table, find as accurately as the data will permit, l_x for values of $x=35$, 42 and 47 given $l_{20}=512$, $l_{30}=439$, $l_{40}=346$, $l_{50}=243$.

[I.A.S. 1948] (Ans. 394, 326, 274)

2. Estimate by Newton's method of interpolation, the expectation of life at age 22 from the following data, stating the assumption underlying the formula used by you—

Age	10	15	20	25	30	35
Expectation of life (in yrs.)	35·4	32·2	29·1	26·0	23·1	20·4

[I.A.S. 1949, 1965] (Ans. 27·85 yrs.)

3. From the following table, find the number of students who obtained less than 45 marks :

Marks	*No. of Students*	*Marks*	*No. of Students*
30—40	31	60—70	35
40—50	42	70—80	31
50—60	51		

[I.A.S. 1967 ; I.C.W.A. Jan. 1965] (Ans. 48 students)

4. The wages earned by workers per month in a certain factory are given at the next page. Calculate the number of workers earning more than Rs. 75 per month.

Monthly income	*No. of workers*
up to Rs. 50	50
„ „ „ 60	150
„ „ „ 70	300
„ „ „ 80	500
„ „ „ 90	700
„ „ „ 100	800

[B.Com. Nagpur 1963] (Ans. 404)

5. From the following data estimate the number of persons in the income group of Rs. 20 to Rs. 25.

Income	*Number of persons*
below Rs. 10	20
„ „ 20	45
„ „ 30	115
„ „ 40	210
„ „ 50	325

[B.Com. Nagpur 1969] (Ans. 31)

6. Find y for $x=2$ from the following table,

x.........	0	1	3	4	5
y.........	39	85	151	264	388

[I.C.W.A. Jan. 1969] (Ans. 96·4)

7. Estimate by suitable method of interpolation the number of persons whose income is Rs. 19 but does not exceed Rs. 25 from the following data :

Income in Rs.	*No. of persons*
1 and not exceeding 10	50
10 „ „ „ 19	70
19 „ „ „ 28	203
28 „ „ „ 37	406
37 „ „ „ 46	304

[M.A. Rajasthan, 1965] (Ans. 107)

8. The following data show the monthly average number of deaths under one year in a certain large city. Find the missing term :

Year	1960	1961	1962	1963	1964
No. of deaths (monthly average)	940	?	907	843	798

[I.C.W.A. Jan. 1972] (Ans. 952)

9. The following values are given in a table :

x	1	2	3	4	5
y	2,16,000	2,26,981	?	2,50,047	2,62,144

Using any suitable algebraic method, find the value of y for $x=3$. [I.A.S. 1953] (Ans. 2, 38, 328)

10. Mantissæ of four numbers are given below :

Numbers	4200	4210	4220	4230
Mantissæ	62,32,493	62,42,821	62,53,125	62,63,404

Find the mantissæ of logarithm of 4,213.

[I.C.W.A. July 1968] (Ans. 62, 45, 915)

11. Mention a formula which will help interpolation when observations are shown to be at unequal intervals.

The observed values of a function are respectively 168, 120, 72 and 63 at the four positions 3, 7, 9 and 10 of the independent varies. What is the last estimate you can give for the value of the function at the position 6 of the independent variable ? (Ans. 147)

12. The following figures relate to the number of estates liable to estate duty in a particular year :

Class of estate	*Number liable*
Rs. 25,000 — Rs. 30,000 ...	638
" 30,000 — " 40,000 ...	740
" 40,000 — " 50,000 ...	415

Estimate the number between Rs. 31,000 and Rs. 32,000 by interpolation. (Ans. 85)

13. Comment on the necessity and usefulness of interpolation. Describe the graphic method of interpolation. [I. C. W. A. Jan. 1970]

14. From the following table find the interpolated figure for the populations in 1946.

years	1930	1940	1950	1960
Population of a town	25,494	29,003	32,528	36,070

[I.C.W.A. 1962] (Ans. 31,116)

15. The population of a country in the decennial census was as under. Estimate the population for 1955 :

years	1921	1931	1941	1951	1961
Populations (in '000)	46	66	81	93	101

[I.C.W.A. 1963] (Ans. 99·56 thousands)

16. The followings are the amounts of income-tax paid by a few businessmen during one year :

more than Rs. 500	600
" " " 1,000	550
" " " 1,500	425
" " " 2,000	275
" " " 2,500	100
" " " 3,000	25

Find out the number of businessmen who paid more than Rs. 1,200 but not more than Rs. 2,400 as income-tax.

[I.C.W.A. July 1965] (Ans. 369)

17. The following table gives the normal weight of babies during the first twelve months of life :

age (months)	0	2	5	8	10	12
weight (lb.)	$7\frac{1}{2}$	$10\frac{1}{4}$	15	16	18	21

Find the weight of a 7 months old baby. (Ans. 15·66 lbs.)

18. From the following table, by using Newton's backward interpolation formula, find the value of y corresponding to $x=38$.

x	10	20	30	40
y	4	9	17	22

(Ans. 21·528)

19. Applying Newton's backward formula, find the value of $x=32$ from the table given below :

x	10	15	20	25	30
y	20	16	22	24	28

(Ans. 34·9816)

20. Find out with the help of E the value of the production for the year 1974 from the following table :

year	1970	1971	1972	1973
production ('000) tons	20	21	24	27

(Ans. 28 in thousand tons)

21. Find $f(x)$ given that $f(0)=-3$, $f(1)=6$, $f(2)=8$, $f(3)=12$ (State your assumption, if any). Hence find $f(6)$.

[I.C.W.A. June 1976] (Ans. 126)

22. Below are given the values of a function U_x for certain values of x :

x :	0	1	2	3	4
U_x :	1	0	5	22	57

Construct the table of differences. What does this table suggest ? Use this table to find U_5. [I.C.W.A. Dec. 76] (Ans. 116)

23. State Lagrange's interpolation formula. Use it to find the value of U_4 of a function U_x, given that $U_1=10$, $U_2=15$, $U_5=42$.

[I.C.W.A. Dec. '76] (Ans. 31)

24. Estimate U_2 from the following table :

X :	1	2	3	4	5
U_x :	7	*	13	21	37

State the necessary assumptions made.

[I.C.W.A. June '77] (Ans. 9·5)

25. The following table gives the expectation of life (e_x^0) at age x. Calculate expectation of life at age 12 by using Newton's forward interpolation formula :

x :	10	15	20	25	30	35
e_x^0 :	35·4	32·2	29·1	26·0	23·1	20·4

[I.C.W.A. Dec. '77] (Ans. 34·174)

26. Discuss the difference between Newton's forward and backward interpolation formulæ. Given,

$\log_{10}654=2·8156$, $\log_{10}658=2·8182$, $\log_{10}659=2·8189$, $\log_{10}661=2·8202$. —Find by Lagrange's interpolation formula $\log_{10}656$ (retain four decimal places in your answer).

[I.C.W.A. June '78] (Ans. 2·8168)

27. Explain, what do you understand by the symbolic operator E. Given the following table, construct a difference table and from it estimate y when $x=0{\cdot}35$ by using Newton's backward interpolation formula.

x:	0	0·1	0·2	0·3	0·4
y:	1	1·095	1·179	1·251	1·310

(Answer to be given correct to 3 dec. places)
[I. C. W. A. June '78] (Ans. 1·282)

28. Apply the appropriate interpolation formula to find log 3·146 given log 3·141 = 0·4970679, log 3·142 = 0·4972062, log 3·143 = 0·4973444, log 3·144 = 0·4974825, log 3·145 = 0·4976205 (Find correct up to seven decimal places.) [I. C. W. A. Dec. '78] (Ans. 0·4977584)

29. State Lagrange's interpolation formula. The mode of a certain frequency curve $y=f(x)$ is attained at $x=9{\cdot}1$ and the value of the frequency function $f(x)$ for $x=8{\cdot}9$, 9·0 and 9·3 are respectively equal to 0·30, 0·35 and 0·25. Calculate the approximate value of $f(x)$ at the mode. [I. C. W. A. Dec. '78] (Ans. 0·36)

30. Given $f(45)=0{\cdot}7071$, $f(50)=0{\cdot}7660$, $f(55)=0{\cdot}8192$ and $f(60)=0{\cdot}8693$, find $f(59)$ correct to 4 places of decimal.
[I. C. W. A. June 79] (Ans. 0·8683)

31. Find with the help of the symbolic operator E the value of $\log_{10}666$ from the following table :

$\log_{10}654=2{\cdot}8156$, $\log_{10}658=2{\cdot}8182$, $\log_{10}662=2{\cdot}8209$.

[I. C. W. A. June '79] (Ans. 2·8237)

32. The values of a function $f(x)$ are given below for some specified values of x :

x:	3	4	5	9
$f(x)$:	6	5	−2	30

Using an appropriate interpolation formula, find the value of $f(7)$.
[C. U. B. Com. (Hons.) 1980] (Ans. −10)

33. Find the missing term in the following table :

x:	0	1	2	3	4
y:	1	3	9	*	81

[I. C. W. A. June 1980] (Ans. 31)

6

FREQUENCY DISTRIBUTION

Observation, Frequency

Suppose the weekly wages (in Rs.) of 60 workers, in a certain factory, are collected by an investigator either from the office-records or by personal interviews.

The collected data are as follows (*in Rs.*) :

45, 20, 50, 10, 25, 28, 17, 28, 45, 30, 9, 18, 37, 45, 32,
45, 41, 35, 37, 45, 36, 32, 40, 37, 40, 45, 32, 45, 17, 17,
9, 18, 28, 35, 32, 47, 20, 25, 28, 26, 25, 17, 19, 30, 35,
32, 26, 21, 26, 20, 30, 10, 10, 40, 20, 28, 50, 50, 30, 40.

Here the variable observed is the 'weekly wages' and the data obtained are the observations or observed values.

The raw data recorded, appear in a complex and arbitrary manner. One cannot fully grasp the true significance of the figures, at a first sight. So some modifications are necessary. Therefore, the data should be arranged in a definite order, either ascending or descending.

Here the above data are arranged in ascending order which are as follows—

9, 9, 10, 10, 10, 17, 17, 17, 17, 18, 18, 19, 20, 20, 20,
20, 25, 25, 25, 26, 26, 26, 28, 28, 28, 28, 28, 30, 30, 30,
30, 32, 32, 32, 32, 32, 35, 35, 35, 35, 37, 37, 37, 40, 40,
40, 40, 41, 41, 45, 45, 45, 45, 45, 45, 45, 47, 50, 50, 50.

The 60 observations are not all different, some of them are repeated. The distinct observations are known as the *values of the variable.*

The above arrangement can also be represented in the form of a table as shown below—

TABLE I

Weekly wages (Rs.)	*No. of workers*	*Weekly wages* (Rs.)	*No. of workers*
9	2	30	4
10	3	32	5
17	4	35	4
18	2	37	3
19	1	40	4
20	4	41	2
25	3	45	7
26	3	47	1
28	5	50	3

A characteristic which can be expressed numerically is called a **variate** or **variable.** The number of times each variate occurs is known as its **frequency.** A frequency table is a chart consisting of the variates with their respective frequencies. A classification showing different values of a variate and the corresponding frequency is known as **frequency distributions.**

In the above table, weekly wages are the variates and the number of workers getting the same wage is the frequency. Here 9 occurs 2 times, 10 occurs 3 times, etc. Frequencies of the values (of variate) 9, 10, ..., etc. are respectively 2, 3, ..., etc.

Frequency Distribution, Simple and Grouped.

There are two types of frequency distribution :

(1) Simple frequency distribution ;

(2) Grouped frequency distribution.

(1) *Simple frequency distribution.*

This shows the values of the variate individually. *For exmple,* Table I (shown above).

(2) *Grouped frequency distribution.*

The above arrangement of data (shown in Table I) is suitable for a small number of figures. Now suppose there are a huge number of figures, say, 1,000, then the above method of arrangement will not be helpful to the statistician for application of any kind of mathematical principles.

In such cases, the values of the variate may be shown in groups or intervals giving rise to a grouped frequency distribution, as shown below—

TABLE II : *Grouped Frequency Distribution*

Variate (weekly wages) (in Rs.)	*Frequency* (no. of workers)
from 1 to 10	5
" 11 " 20	11
" 21 " 30	15
" 31 " 40	16
" 41 " 50	13
Total	60

Few Terms (*Associated with grouped frequency distribution*) :

(*a*) Class-interval.

(*b*) Class-frequency, total frequency.

(*c*) Class-limits (upper and lower).

(*d*) Class-boundaries (upper and lower).

(*e*) Mid-value of class-interval.

(*f*) Width of class-interval.

(*g*) Frequency density.

(*h*) Percentage frequency.

(*a*) *Class-interval.*

A large number of observations having a wide range, is usually classified in several groups. Each of these groups is known as class-interval (or class). In Table II, class-intervals are 1—10, 11—20, ..., etc. In all there are 5 class-intervals, of which the first class-interval is 1—10, and the last class-interval is 41—50.

If one end of a class-interval is not given, then it is known as an open-end class. There may be two open-end classes. For example, less than 15, 15—20, 20—25, above 25. The class-interval having zero frequency is known as *empty class.*

(*b*) *Class-frequency, Total frequency.*

The number of observations (frequency) in a particular class-interval is known as *class-frequency.* In Table II, for the class-interval 1—10, class-frequency is 5 ; for the class-interval 11—20, class-frequency is 11 and so on. The sum of all class-frequencies is called the total frequency. In the Table, it is 60. Total frequency is the total number of observations.

(*c*) *Class-limits.*

The two ends of a class-interval are called class-limits. Of a particular class-interval, the smaller and greater numbers are known as *lower* and *upper* class-limits respectively. In Table II, for the first class, class-limits are 1 and 10, whereas lower class-limit is 1 and upper class-limit is 10. For the next class, lower and upper class-limits are 11 and 20 respectively.

(*d*) *Class-boundaries.*

The class-boundaries may be calculated from the class-limits by the following rule :

lower class-boundary = lower class-limit $-\frac{1}{2}d$,

upper class-boundary = upper class-limit $+\frac{1}{2}d$,

where d = common difference between upper class of any class-interval with the lower class of the next class-interval.

In Table II, $d=1$, for the first class-interval,

lower class-boundary $=1-\frac{1}{2}\times 1=1-{\cdot}5=0{\cdot}5$,

upper class-boundary $=10+\frac{1}{2}\times 1=10+{\cdot}5=10{\cdot}5$.

Again, for the next class-interval, lower class-boundary = 10·5

and upper class-boundary = 20·5 and so on.

Example.

Find the class-boundaries of (i) $20-24{\cdot}9$, $25-29{\cdot}9$, ... (ii) $20-25$, $25-30$, ...

(i) $d={\cdot}1$, class-boundaries are, $19{\cdot}95-24{\cdot}95$, $24{\cdot}95-29{\cdot}95$, ...

(ii) $d=0$, class-boundaries are, $20-25$, $25-30$, ...

(e) *Mid-value.*

The value exactly at the middle of a class-interval is known as its *mid-value.* It is calculated by adding the two class-limits divided by 2 (or, adding the two class-boundaries divided by 2).

In Table II, for the first class-interval, mid-value $=\dfrac{1+10}{2}=5{\cdot}5$,

for the second class-interval, mid-value $=\dfrac{11+20}{2}=15{\cdot}5$ and so on.

(f) *Width.*

The width (or size) of a class-interval is the difference between the class-boundaries (not class-limits)

∴ width = upper class-boundary − lower class-boundary.

In Table II, for the first class, width $=10{\cdot}5-{\cdot}5=10$,

for the second class, width $=20{\cdot}5-10{\cdot}5=10$, and so on.

(g) *Frequency density.*

It is the ratio of the class-frequency to the width of that class-interval, *i.e.*,

$$\text{Frequency density}=\frac{\text{class-frequency}}{\text{width of the class}}.$$

In Table II, for the first class, frequency density $=\dfrac{5}{10}={\cdot}5$,

for the second class, frequency density $=\dfrac{11}{10}=1{\cdot}1$ and so on.

(h) *Percentage frequency.*

It is the ratio of class-frequency to total frequency expressed as percentage, *i.e.*,

$$\text{Percentage frequency}=\frac{\text{class-frequency}}{\text{total frequency}}\times 100$$

$$(=\text{Relative frequency}\times 100).$$

In Table II, for the class-frequency 5, % frequency $=\dfrac{5}{60}\times 100$

$=8{\cdot}33$

... ... 11, % frequency $=\dfrac{11}{60}\times 100=18{\cdot}33$ and so on.

Now all these terms are illustrated in the following table, with reference to Table II.

TABLE III

Illustration of Class-limits, Class-boundaries, Mid-value, Width, etc.

(Data : Table II)

Class-interval	Class-frequency	*Class-limits*		*Class-boundaries*		Mid-value	Width of class	Frequency density	Percentage frequency
		lower	upper	lower	upper				
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1—10	5	1	10	0·5	10·5	5·5	10	0·5	8·33
11—20	11	11	20	10·5	20·5	15·5	10	1·1	18·33
21—30	15	21	30	20·5	30·5	25·5	10	1·5	25·00
31—40	16	31	40	30·5	40·5	35·5	10	1·6	26·67
41—50	13	41	50	40·5	50·5	45·5	10	1·3	21·67
Total	60	—	—	—	—	—	—	—	100·00

Discrete and Continuous Series.

In Table II, the first group ends at Rs. 10, while the second group starts at Rs. 11, and there is a gap of Re. 1, between these groups. Similar gap lies between any other two groups. So there is a discontinuity of the series, usually known as *discrete series.* By slight alteration of the groups, a discrete series may be converted into a continuous series. Instead of writing the first group as Re. 1 to Rs. 10, write as Re. 1 and less than Rs. 11, similarly for the second group, Rs. 11 and less than Rs. 21, etc.

Hence Table II, can be written as a Continuous Series as follows :

TABLE IV : Continuous Series (Data : Table II)

Frequency distribution of weekly wages obtained by 60 workers

Weekly wages (Rs.) (variate)	*frequency*
1 and less than 11	5
11 21	11
21 31	15
31 41	16
41 51	13
Total	60

Note. Discrete and continuous series are also discussed in the chapter of Classification.

Compilation of Frequency Distribution Table from Raw Data.

Instead of going through so many stages, we can directly get the Table II, from the raw data collected, with the help of *Tally Marks.*

In counting values, a vertical line (/) is used for one, four vertical lines crossed by a diagonal line (𝍸) are used for every five countings belonging to the same group, as shown below :

TABLE V : *Frequency distribution of weekly wages*

Weekly wages (Rs.)	*Tally marks*	*Frequency*
1—10	𝍸	5
11—20	𝍸 𝍸 𝍷	11
21—30	𝍸 𝍸 𝍸	15
31—40	𝍸 𝍸 𝍸 𝍷	16
41—50	𝍸 𝍸 𝍷𝍷𝍷	13
Total	—	60

Construction of Frequency Distribution Table.

The process of constructing a frequency distribution table from raw data are as follows—

1. The smallest and largest figure are to be located first and then find the range (*i.e.*, the difference between the largest and smallest figures).

2. Distribute the range in a suitable number of class-intervals. The number of these class-intervals should preferably be between 6 and 15, depending on the number of observations. However, there is no rigidity about it. The class-intervals may be more than 15, depending upon the total number of observations.

The class-limits should be chosen in such a way, that most of the observations lie within the class-limits. Same width of the class-intervals are preferred. Of course, there may be unequal widths in some cases.

3. The number of observations (*i.e.*, frequency) lying in each class-interval is determined by Tally Marks.

4. The construction of a table will be complete, by placing class-intervals in the first column, tally-markings in the second column, and corresponding class-frequencies in the third column.

Note. *Choice of number of class-intervals.* The number of class-intervals should neither be too large (as frequency distribution will be very large) nor too small (as essential characteristics of the distribution will not be revealed). As a working rule, the number lies between 6 and 15. For lesser number of observations, Sturge's formula may also be used, $n = 1 + 3{\cdot}3 \log (N)$, where n = the number of classes, N = total frequency. To ensure continuity and to get correct class-interval, we should adopt *exclusive* method of classifications.

Choice of Class-limits.

The lower limit of the first class-interval should either 0 or 5 or multiple of 5. For example, if lowest data is 23, then for a width of class-interval 5, the first class should be 20—25 (instead of 19—24). The approximate width of the classes may be obtained by dividing the range of the data by the number of classes. Further the class-limits should be so chosen that the observations occurring most frequently lie within the class-limits preferably near the mid-value of the class-limits.

Example.

The following is an array of 65 marks obtained by students in a certain examination :

26,	45,	27,	50,	45,	32,	36,	41,	31,	41,	48,	27,	46,
47,	31,	34,	42,	45,	31,	28,	27,	49,	48,	47,	32,	33,
35,	37,	47,	28,	46,	26,	46,	31,	35,	33,	42,	31,	41,
45,	42,	44,	41,	36,	37,	39.	51,	54,	53,	38,	55,	39,
52,	38,	54,	36,	37,	38,	56,	59,	61,	65,	64,	72,	64.

Draw up a frequency distribution table classified on the basis of marks with class-intervals of 5.

TABLE VI : *Tally Sheet*

Class-intervals of marks	*Tally marks*	*Frequency*
25—29	卌 //	7
30—34	卌 卌	10
35—39	卌 卌 ///	13
40—44	卌 ///	8
45—49	卌 卌 ///	13
50—54	卌 /	6
55—59	///	3
60—64	///	3
65—69	/	1
70—74	/	1
Total	—	65

Now the required frequency distribution is shown below :

TABLE VII : *Frequency distribution of marks obtained by 65 students*

Marks	*Frequency*
25—29	7
30—34	10
35—39	13
40—44	8
45—49	13
50—54	6
55—59	3
60—64	3
65—69	1
70—74	1
Total	65

Cumulative Frequency Distribution.

It is a form of frequency distribution in which each frequency beginning with the second from the top is added with the total of the previous ones, the class-intervals being adjusted accordingly.

Example.

TABLE VIII : *Cumulative frequency distribution showing the marks.* (Data : reference Table VII)

Marks	*Frequency*	*Cumulative frequency*
25—29	7	7
30—34	10	17
35—39	13	30
40—44	8	38
45—49	13	51
50—54	6	57
55—59	3	60
60—64	3	63
65—69	1	64
70—74	1	65
Total	65	—

The cumulative frequency up to 34 (up to the upper class-limit of the second class-interval) is obtained by adding the frequency of the second class with that of the previous class and so on. This kind of cumulative frequency is known as *'less than type'* cumulative frequency when addition is done from top. Conversely if the addition is done from below, then it will be *'greater than type'* cumulative frequency.

It may be **noted** that for less than cumulative frequency corresponding to the highest class-boundary and greater than cumulative frequency corresponding to the lowest class-boundary must be equal to the total frequency. Also the sum of these two types of cumulative frequencies at any stage of variate is the total frequency.

Uses of Cumulative Frequency.

It is used (*a*) to find the number of observations less than or greater than any given variate ; (*b*) to find the number of observation lying between any particular class-interval ; (*c*) to find median, quartiles, deciles and percentiles graphically (will be discussed in the chapter of *Average*).

Example.

From the following table find (*a*) the less than and (*b*) greater than cumulative frequencies, (*c*) cumulative frequency distribution, (*d*) cumulative percentage distribution.

Wages (Rs.)	11—20	21—30	31—40	41—50	51—60	61—70	Total
Frequency	5	7	12	15	8	3	50

The class-boundaries of the class-intervals are respectively 10·5—20·5, 20·5—30·5, ..., etc. (ref. Table III). The boundary points are 10·5, 20·5, 30·5, ..., etc. There is no frequency below 10·5, so its cumulative frequency is 0 ; the frequency below 20·5 is 5, the frequency below 30·5 is 12 (=5+7), the frequency below 40·5 is 24 (=12+12) and so on. This is less than cumulative frequency. For greater than type, we are to start adding from the end. Now corresponding to the class-boundaries 70·5, 60·5, 50·5, 40·5, ···, etc. the respective cumulative frequencies are 0, 3, 11 (=3+8), 26 (=11+15), ..., etc.

Cumulative frequency distribution consists of the variates (class-boundary points only) with the corresponding less than cumulative frequencies.

TABLE IX : *Cumulative Frequencies*

Wages (Rs.)	*Cumulative frequencies* (less than)	*Cumulative frequencies* (greater than)
10'5	0	50
20'5	5	45
30'5	12	38
40'5	24	26
50'5	39	11
60'5	47	3
70'5	50	0

Cumulative frequency and Cumulative percentage distributions

Wages (Rs.)	*Cumulative frequency* (less than)	*Cumulative percentage* (less than)
10'5	0	0
20'5	5	10
30'5	12	24
40'5	24	48
50'5	39	78
60'5	47	94
70'5	50	100

Example.

Present the following data of the percentage marks of 60 students in the form of a Frequency Table with 10 classes of equal width, one class being 40—49.

41	17	83	63	54	92	60	58	70	06
67	82	33	44	57	49	34	73	54	63
36	52	32	75	60	33	09	79	28	30
42	93	43	80	03	32	57	67	24	64
63	11	35	82	10	23	00	41	60	32
72	53	92	88	62	55	60	33	40	57

[C. A. 1966]

Here minimum marks is 00, maximum mark is 93, and width of class-interval is 10 (from given the class 40—49).

Frequency distribution of marks obtained by 60 students

Marks	*Frequency*
00—09	4
10—19	3
20—29	3
30—39	10
40—49	7
50—59	9
60—69	11
70—79	5
80—89	5
90—99	3
Total	60

Example.

Age at death of 50 persons of a town are given below—

36	48	50	45	49	31	50	48	43	42
37	32	40	39	41	47	45	39	43	47
38	39	37	40	32	52	56	31	54	36
51	46	41	55	58	31	42	53	32	44
53	36	60	59	41	53	58	36	38	60

(*a*) Arrange the data in a frequency distribution in 10 class-intervals ; and

(*b*) Obtain the percentage frequency in each class.

[B. Com. C. U. 1972]

Here the minimum and maximum observations are respectively 31 and 60, and their difference is 29. For 10 classes, width of the classes should be $\frac{29}{10} = 2{\cdot}9$ or 3. Taking 3 as width of the classes, Frequency Distribution Table is drawn.

Frequency Distribution Table and Percentage frequency

Age	*Frequency*	*Percentage frequency*
31—33	6	12
34—36	4	8
37—39	7	14
40—42	7	14
43—45	5	10
46—48	5	10
49—51	4	8
52—54	5	10
55—57	2	4
58—60	5	10
Total	50	100

Example.

If the class mid-points in a frequency distribution of weights of students are 128, 137, 146, 155, 164, 173 and 182 pounds, find (*a*) the class-interval size, (*b*) the class-limits. [C. A. May 1964]

The difference between any two consecutive terms of the class mid-points is 9. Taking class-boundaries, the class-limits of the first class being 123·5 (=128−4·5) and 132·5 (=128+4·5). In the same way, the other class-limits will be 132·5−141·5, 141·5−150·5, 150·5−159·5, 159·5−168·5, 168·5−177·5, 177·5−186·5.

EXERCISE 5

1. Discuss the various steps in the preparation of frequency distribution from raw data. [C. U. M. Com. 1969]

2. Discuss the problems in the construction of a frequency distribution from raw data, with particular reference to the choice of number of classes and class-limits. [I. C. W. A. Jan. 1972]

3. What do you mean by a cumulative frequency distribution ? Point out its special advantages and uses. [I. C. W. A. Jan. 1971]

4. Explain the terms : class-interval, class-limits, class mid-point and class-frequency. [C. A. Nov. 1964]

5. Taking the class-limits 5 – 9, 10 – 14, 15 – 19, ⋯, etc., construct a frequency distribution of the following sets of observations :—

7	27	10	19	39	24	24	24	41	20
23	44	47	36	53	20	16	45	23	22
10	13	31	11	30	21	31	22	28	17
27	32	42	20	15	34	21	29	44	21
59	36	22	18	27	23	21	25	17	28
34	23	48	32	49	29	21	52	43	40
33	37	21	40	22	14	38	28	23	25
27	16	29	20	17	23	19	23	45	35
22	33	15	18	25	38	24	22	13	27
12	24	19	9	12	24	30	35	37	22

(*Ans.* 2, 8, 12, 30, 14, 10, 9, 7, 5, 2, 1) [B.A. (Hons.) C.U. 1964]

6. From the following observation, prepare a Frequency Distribution Table in ascending order starting with 5 – 10 (exclusive method) marks in English :

12	36	40	30	28	20	19	10	10	16
19	27	15	26	20	19	7	35	33	21
26	37	5	20	11	17	37	30	20	5

(*Ans.* frequencies 3, 4, 6, 5, 4, 3, 4). [B. Com., Bangalore, 1968]

7. You are given below the wages paid to some workers in a small factory. Form a frequency distribution with class-interval 10 paise.

(*Wages in Rs.*)

1·10	1·13	1·44	1·44	1·27	1·17	1·98	1·36	1·30
1·27	1·24	1·73	1·51	1·12	1·42	1·03	1·58	1·46
1·40	1·21	1·62	1·31	1·55	1·33	1·04	1·48	1·20
1·60	1·70	1·09	1·49	1·86	1·95	1·50	1·82	1·42
1·29	1·54	1·38	1·87	1·41	1·77	1·15	1·57	1·07
1·65	1·36	1·67	1·41	1·55	1·22	1·69	1·67	1·34
1·45	1·39	1·25	1·26	1·75	1·57	1·53	1·37	1·59
1·19	1·52	1·56	1·32	1·81	1·40	1·47	1·38	1·62
1·76	1·28	1·92	1·46	1·46	1·35	1·16	1·42	1·78
1·68	1·47	1·37	1·35	1·47	1·43	1·66	1·56	1·48

(*Ans.* class-intervals : 1·01 – 1·10, 1·10 – 1·20, ..., etc.

frequencies : 5, 7, 10, 15, 18, 14, 9, 5, 4, 3) [C.A. May 1967]

8. Construct a frequency distribution showing the frequencies with which words of different number of letters occur in the extract reproduced below (omitting punctuation marks), treating as the variable the number of letters in each word :

'A candidate at the time of applying for registration as a student of the institute should be not less than eighteen years of age and have passed the Intermediate Examination of a University constituted by law in India or an examination recognized by the Central Government as equivalent thereto, or the National Diploma in Commerce Examination or the diploma in Rural Services Examination conducted by the National Council of Rural Higher Education.'

[I. C. W. A. July 1967]

(*Ans.* Number of letters

in a word	1	2	3	4	5	6	7	8	9	10	11	12	total
frequency	3	19	12	4	4	3	6	6	4	4	5	2	72

)

9. The following table gives the scholastic aptitude scores of the 50 departmental students of a certain department in a certain university :

345	530	556	354	590
395	515	479	494	420
563	444	629	440	485
505	604	490	445	605
402	406	730	506	516
472	475	610	586	523
691	520	465	468	545
624	582	570	578	505
523	575	420	605	527
461	440	585	420	384

(i) Construct a frequency distribution table with appropriate class-limits and class-boundaries (take the length of the class equal to 30 units).

(ii) Draw histogram to represent the above frequency distribution. [I. C. W. A. Dec. 1978]

(*Ans.* (i) class-limits : 345 – 374, 375 – 404, ...
class-boundaries : 344·5 – 374·5, 374·5 – 404·5, ...
frequency : 2, 3, 4, 5, 8, 8, 3, 6, 7, 2, 0, 1, 1
(ii) using the class-boundaries, histogram is to be drawn.)

10. The weights in Kilogram of 50 persons are given below :

76	64	53	55	66	72	52	63	46	51
53	56	65	60	47	55	67	73	44	54
64	74	48	59	72	61	43	69	61	58
42	52	62	72	43	63	71	64	58	67
46	55	65	75	48	59	67	77	64	78

Arrange the above data in a frequency distribution with class-interval of 5 kg. Construct the frequency polygon on a graph paper with above data. [C.U. B. Com. (Pass) 1980]

11. Marks obtained by 50 students in a History paper of full marks 100 are as follows :

78	25	25	40	30	29	35	42	43	43
44	20	48	44	43	48	36	46	48	47
36	60	31	47	33	65	68	73	39	12
60	20	47	49	51	38	49	35	52	61
34	76	79	20	16	70	65	39	60	45

Arrange the data in a Frequency Distribution Table in class-intervals of length 5 units. Draw a histogram to present the above data. [I.C.W.A. June 1980]

12. The marks scored by 50 students in Geography are as follows :

30	45	48	55	39	25	31	12	18	21
54	59	51	33	43	44	10	38	19	26
41	35	37	41	46	33	51	37	58	48
17	19	23	26	29	38	57	36	35	44
43	27	31	43	22	31	47	34	18	15

Prepare a Frequency Table with 5 class-intervals each of width 10 marks and hence draw the ogive of both types.

7

AVERAGE

Introduction.

We have seen in the earlier chapter how statistical data are condensed to a large extent by tabulation. For practical purpose, tabulation is not enough, especially when we are to compare two or more series of data. In order to make them comparable, it is essential to reduce the figures into one figure. For example, it is required to compare the daily wages obtained by 100 workers belonging to a factory *A* with the daily wages of 100 workers of factory *B*. It would be impossible to arrive at any conclusion, if these two series are directly compared. Now, if each of these series is represented by one figure, comparison would be extremely easy affair.

Satistical data when represented by graphs or diagrams appeal, more to the eye than to the mind. Unless we are able to describe the main theme of a series or what it tends to suggest, we cannot deal with the series adequately. There is a necessity for some single measurement which may give the summary description of the characteristics of a large group of variables.

In most frequency distributions, we find that the tabulated values show small frequencies as the class-limits, while at the middle part frequency is highest. This indicates that near the central part of the distribution, most of the items of the series cluster. Such figures are known as *Measures of Central Tendency* or *Averages*. Average represents a whole series, as such its value lies between the minimum and maximum values and generally it is located in the centre of the distribution.

The *object* of an average is to represent a number of variates in a *simple* and *concise manner*. So it is a representative figure of the entire data. Secondly, it is a basis of comparison with other groups.

Types of Average.

Broadly speaking, there are three types of Measures of Central Tendency (or Avarage) :

1. MEAN 2. MEDIAN 3. MODE

Mean, again, is divided into three types :

1. Arithmetic Mean (A.M.)
2. Geometric Mean (G.M.)
3. Harmonic Mean (H.M.)

Unless specially mentioned, the term *mean* generally refers to the arithmetic mean. It has the maximum application amongst these three types. Median and Mode are known as *Average of Position*, while Mean is known as *Mathematical Average.*

Use of Sigma (Σ) Notation.

The sum $x_1+x_2+x_3+\cdots+x_n$ is often denoted by $\sum\limits_{i=1}^{n} x_i$ or Σx.

Similarly, the sum $p_1x_1+p_2x_2+p_3x_3+\cdots\cdots+p_nx_n$ may be denoted by $\sum\limits_{i=1}^{n} p_i x_i$ or Σpx

The quotient $\dfrac{p_1x_1+p_2x_2+\cdots+p_nx_n}{p_1+p_2+\cdots+p_n}$ is denoted by $\dfrac{\Sigma px}{\Sigma p}$.

Arithmetic Mean (*A.M.*).

Definition : The Arithemetic Mean of the values of a variate $x_1, x_2, x_3, \cdots, x_n$ is the sum of the values divided by their number. Now if $\bar{X}$ denotes the A.M. of the quantities, then

$$\bar{X}=\frac{x_1+x_2+x_3+\cdots+x_n}{n}=\frac{\Sigma x}{n}.$$

This mean (A.M.) is known as *Simple Arithmetic Mean.*

Example.

To find the A. M. of the numbers 2, 5, 9, 11, 13.

Here n (the total number of items) $=5$.

Now, A. M. $(\bar{X})=\dfrac{2+5+9+11+13}{5}=\dfrac{40}{5}=8.$

Weighted Arithmetic Mean (*Weighted Mean*).

Definition : It the n values of a variate $x_1, x_2, x_3, \ldots, x_n$ are taken $f_1, f_2, f_3, \ldots, f_n$ times, respectively (*i.e.*, if, $f_1, f_2, f_3, \ldots, f_n$ are the respective frequencies of $x_1, x_2, x_3, \ldots, x_n$) then

$$\text{Weighted Mean } (\bar{X})=\frac{f_1x_1+f_2x_2+f_3x_3+\cdots+f_nx_n}{f_1+f_2+f_3+\cdots+f_n}=\frac{\Sigma fx}{\Sigma f}.$$

Example.

To find average income from the following table :

Daily income (Rs.)	2	5	9	11	13	Total
No. of workers	2	1	4	2	1	10

Variate (x) (income in Rs.)	*Frequency* (f) (no. of workers)	fx
(1)	(2)	(3) = (1) × (2)
2	2	4
5	1	5
9	4	36
11	2	22
13	1	13
Total	10	80

$\therefore$ Weighted Mean (or average income) $= \dfrac{\Sigma fx}{\Sigma f} = \dfrac{80}{10} =$ Rs. 8·00

Note. Here, $x_1=2$, $x_2=5$, $x_3=9$, $x_4=11$, $x_5=13$ and $f_1=2$, $f_2=1$, $f_3=4$, $f_4=2$, $f_5=1$.

This method of computing A. M. is known as Direct Method.

Important Property of Arithmetic Mean.

(1) *The algebraic sum of the deviations of the values from their arithmetic mean is zero.*

Proof. (*a*) (*Simple*) *Arithmetic Mean* :

The differences $x_1-\bar{X}$, $x_2-\bar{X}$, $x_3-\bar{X}$, ... , $x_n-\bar{X}$ (irrespective of sign) are called the *deviations* of $x_1, x_2, x_3, \ldots, x_n$ respectively from the mean $\bar{X}$. Now,

$$\sum_{i=1}^{n}(x_i-\bar{X})=(x_1-\bar{X})+(x_2-\bar{X})+(x_3-\bar{X})+\cdots+(x_n-\bar{X})$$
$$=(x_1+x_2+\cdots+x_n)-(\bar{X}+\bar{X}+\cdots n \text{ times})$$
$$=\Sigma x-n\bar{X}=n\bar{X}-n\bar{X}\quad(\bar{X}=\Sigma x/n)$$
$$=0$$

(b) *Weighted Arithmetic Mean :* [C. U. B. Com. (Hons.) 1980]

$$\sum_{i=1}^{n} f_i(x_i-\bar{X})=f_1(x_1-\bar{X})+f_2(x_2-\bar{X})+\cdots+f_n(x_n-\bar{X})$$

$$=f_1x_1+f_2x_2+\cdots+f_nx_n-(f_1\bar{X}+f_2\bar{X}+\cdots+f_n\bar{X})$$

$$=\Sigma fx-(f_1+f_2+\cdots+f_n)\bar{X}$$

$$=\Sigma f\bar{X}-\Sigma f\bar{X}=0 \text{ (as } \bar{X}=\Sigma fx/\Sigma f)$$

Example.

A. M. of 2, 5, 9, 11, 13 is 8.

Now, the deviations are : (2 − 8), (5 − 8), (9 − 8), (11 − 8), (13 − 8), *i.e.*, − 6, − 3, 1, 3, 5 whose sum is $-6-3+1+3+5=-9+9=0$.

Example.

Weighted A. M. of 2, 5, 9, 11, 13 having frequencies 2, 2, 3, 2, 1 is 7·6 (*by the process shown before*).

Deviations are : (2 − 7·6), (5 − 7·6), (9 − 7·6), (11 − 7·6), (13 − 7·6), *i.e.*, − 5·6, − 2·6, 1·4, 3·4, 5·4.

Again $2(-5{\cdot}6)+2(-2{\cdot}6)+3(1{\cdot}4)+2(3{\cdot}4)+1(5{\cdot}4)$

$$=-11{\cdot}2-5{\cdot}2+4{\cdot}2+6{\cdot}8+5{\cdot}4 \qquad =-16{\cdot}4+16{\cdot}4=0.$$

(2) *Prove that for a given set of observations the sum of the squares of deviations is the minimum, when deviations are taken from the arithmetic mean.* [I. C. W. A. Jan. 1971]

$x_i-\text{A}=(x_i-\bar{X})+(\bar{X}-\text{A})$, where x_i are n observations,
$\bar{X}$ is actual mean,
A is any arbitrary constant.

Now $\Sigma(x_i-\text{A})=\Sigma(x_i-\bar{X})+\Sigma(\bar{X}-\text{A})$

And $\Sigma(x_i-\text{A})^2=\Sigma(x_i-\bar{X})^2+\Sigma(\bar{X}-\text{A})^2+2\Sigma(x_i-\bar{X})(\bar{X}-\text{A})$

$$=\Sigma(x_i-\bar{X})^2+n(\bar{X}-\text{A})^2+2(\bar{X}-\text{A})\,\Sigma(x_i-\bar{X})$$

as $(\bar{X}-\text{A})$ is constant.

$$=\Sigma(x_i-\bar{X})^2+n(\bar{X}-\text{A})^2 \quad [\text{as } \Sigma(x_i-\bar{X})=0$$

by prop. (1)].

Now $\Sigma(x_i-\text{A})^2>\Sigma(x_i-\bar{X})^2$ as $n(\bar{X}-\text{A})^2$ is always positive.

The sign of equality holds when and only when $\bar{X}=\text{A}$.

Example.

x	$d=x-\bar{X}$	d^2
2	-4	16
4	-2	4
6	0	0
8	2	4
10	4	16
Total	—	40

$$\text{A. M. }(\bar{X})=\frac{2+4+6+8+10}{5}=6.$$

Now if the deviations are taken from any other value then the sum of the squares of deviations would be greater than 40.

(3) *If $\overline{X}_1, \overline{X}_2$ be the means of two groups having observations N_1, N_2 respectively, then the mean $(\overline{X})$ of the composite group $N(=N_1+N_2)$ is given by the relation* $N\overline{X}=N_1\overline{X}_1+N_1\overline{X}_2$.

This may be generalised for any number of groups

$$N\overline{X}=N_1\overline{X}_1+N_2\overline{X}_2+N_3\overline{X}_3+\cdots.$$

Note. This property has been discussed with illustrations under the heading of *Mean Composite Group* later on.

Short-cut Method of Determining Mean (*Method of Assumed Means*).

(1) *Simple A.M.* :

If d_i $(i=1, 2, \ldots, n)$ be the deviations of the n observations $x_1, x_2, \ldots, x_n$ from any arbitrary value A (as near as possible to the true mean)

then $d_i=x_i-\text{A}$ or $x_i=\text{A}+d_i$

$$\text{Now } \bar{X}=\frac{x_1+x_2+\cdots+x_n}{n}=\frac{(\text{A}+d_1)+(\text{A}+d_2)+\cdots+(\text{A}+d_n)}{n}$$

$$=\frac{n\text{A}+(d_1+d_2+\cdots+d_n)}{n}=\frac{n\text{A}+\Sigma d}{n}=\text{A}+\frac{\Sigma d}{n}$$

Alternatively :

$$x=(x-\text{A})+\text{A}=d+\text{A}$$

$$\text{or}\quad \Sigma x=\Sigma(d+\text{A})=\Sigma d+\Sigma\text{A}=\Sigma d+n\text{A}$$

$$\text{or}\quad \frac{\Sigma x}{n}=\frac{\Sigma d}{n}+\frac{n\text{A}}{n}$$

$$\text{or}\quad \bar{X}=\text{A}+\frac{\Sigma d}{n}$$

Example. To find A.M. of 2, 5, 9, 11, 13 by short-cut method.

Variate x	$d=x-A$
2	-7
5	-4
9	0
11	2
13	4
Total	-5

Let A (assumed mean) = 9

$$\therefore \quad \text{A.M.}=A+\frac{\Sigma d}{n}$$

$$=9+\frac{(-5)}{5}=9-1=8$$

Note. If, it is taken A = 11 or 5, we would get the same result, *i.e.*, if the value of Origin (A) is changed, A.M. remains same.

(2) *Weighted Mean.*

We have, $\overline{X}=\dfrac{f_1x_1+f_2x_2+\ldots+f_nx_n}{N}$, when $\Sigma f=N$

$$=\frac{f_1(A+d_1)+f_2(A+d_2)+\cdots+f_n(A+dn)}{N}$$

$$=\frac{(f_1A+f_2A+\cdots+f_nA)+(f_1d_1+f_2d_2+\cdots+f_nd_n)}{N}$$

$$=\frac{A(f_1+f_2+\cdots+f_n)+\Sigma fd}{N}$$

$$=\frac{AN+\Sigma fd}{N}=A+\frac{\Sigma fd}{N}=A+\frac{\Sigma fd}{\Sigma f}$$

Example.

Variate (x)	*Frequency* (f)	$d=x-A$	fd
(1)	(2)	(3)	(4) = (2) × (3)
2	2	-7	-14
5	2	-4	-8
9	3	0	0
11	2	2	4
13	1	4	4
Total	10	—	-14

Let A (=assumed mean)=9

$$\text{Now A.M.}=A+\frac{\Sigma fd}{\Sigma f}=9+\frac{(-14)}{10}=9-1{\cdot}4=7{\cdot}6.$$

Note. If, it is taken A=11 or 5, we would get the same result. So if the value of origin (A) is changed, mean is unchanged.

Step Deviation Method.

In this method, the only additional point is that we take a common factor (usually the width of class-interval in case of grouped data of equal width or L. C. M. of deviations taken) and multiply the result by the same common factor. This is for simplifying the calculation.

The formula stands :

$$\text{A.M. }(\overline{X})=A+\frac{\Sigma fd}{\Sigma f}\times i,$$ When A=assumed mean, f=frequency,

$$d'=\frac{x-A}{i}=\frac{d}{i},\; i=\text{common factor.}$$

Note. This formula can be proved by putting $d=d'\times i$ in the above formula.

Alternatively :

Let $u_i=\frac{x_i-A}{d}$, where A and d are constants, *i.e.*, variates x_i have been changed to new variates u_i.

Now $x_i=A+du_i$

or $f_ix_i=f_i(A+du_i)$, multiplying both sides by f_i

or $\Sigma f_ix_i=\Sigma f_i(A+du_i)$, taking aggregate (Σ) to both sides

$=\Sigma(f_iA+df_iu_i)=\Sigma f_iA+\Sigma df_iu_i$

$=A\Sigma f_i+d\Sigma f_iu_i$, as A and d are constants

$=AN+d\Sigma f_iu_i$

or, $\frac{\Sigma f_ix_i}{N}=A+d\,\frac{\Sigma f_iu_i}{N}$. dividing by Σf (=N)

or, $\overline{X}=Ad\bar{u}$.

Example.

To find A.M. from the following table :

x:	10	20	30	40	50	60
f:	6	4	6	12	8	4

Calculation of A.M. :

x	(f)	$d=x-A$	$d'=\frac{d}{10}$	fd'	$d_1'=\frac{d}{5}$	fd_1'
(1)	(2)	(3)	(4)=(3)÷10	(5)=(2)×(4)	(6)=(3)÷5	(7)=(2)×(6)
10	6	−30	−3	−18	−6	−36
20	4	−20	−2	−8	−4	−16
30	6	−10	−1	−6	−2	−12
40	12	0	0	0	0	0
50	8	10	1	8	2	16
60	4	20	2	8	4	16
Total	40	—	—	−16	—	−32

Let A (assumed mean) = 40.

In the above Table two common factors (*i.e.*, two scales) 10 and 5 have been taken to scale down the data and shown separately in the Table.

For the scale $i=10$

$$\text{A.M.}=\text{A}+\frac{\Sigma fd'}{\Sigma f}\times i=40+\frac{(-16)}{40}\times 10=40-4=36.$$

Again for the other scale $i=5$

$$\text{A.M.}=\text{A}+\frac{\Sigma fd_1'}{\Sigma f}\times i=40+\frac{(-32)}{40}\times 5=40-4=36.$$

We get the same result for different scales.

Calculation of Mean from Grouped Data (*Continuous Series*).

In case of grouped data, for computing Arithmetic Mean, the only additional work is to find the mid-values of each class-interval. This is done by taking the Arithmetic Mean of class-limits (or class-boundaries). The other steps in the calculations remain the same. The idea will be clear from the example given below. The calculation may be done by applying any one of the following methods.: (i) Direct Method, (ii) Short-cut Method, (iii) Step Deviation Method

Example.

Calculate the mean weight from the following table :

Weight (lbs.)	95—105	105—115	115—125	125—135	Total
No. of students	20	26	38	16	100

Wt. (lb)	*No. of students* (*f*)	*Mid-value* (*x*)	$d=x-\text{A}$	*fd*
95—105	20	$\frac{95+105}{2}=100$	-20	-400
105—115	26	110	-10	-260
115—125	38	120	0	0
125—135	16	130	10	160
Total	100	—	—	-500

Let A = 120

$$\text{A. M.}=\text{A}+\frac{\Sigma fd}{\Sigma f}=120+\frac{-500}{100}=120-5=115 \text{ lbs.}$$

Use of Step Deviation.

Example : Calculate the mean from the following table :

Monthly wages (Rs.) of domestic servants	0—10	10—20	20—30	30—40	40—50
No. of servants	1	4	10	22	30
		50—60	60—70	70—80	80—90
		35	10	7	1

[C.A. Nov. 1962]

Class-interval wages (Rs.)	*Frequency* (f)	*Mid-value* (x)	d $= x - A$	$d' = d + 10$	fd'
0—10	1	5	−50	−5	−5
10—20	4	15	−40	−4	−16
20—30	10	25	−30	−3	−30
30—40	22	35	−20	−2	−44
40—50	30	45	−10	−1	−30
50—60	35	55	0	0	0
60—70	10	65	10	1	10
70—80	7	75	20	2	14
80—90	1	85	30	3	3
Total	120	—	—	—	−98

Let $A = 55$

$$\text{A. M.} = A + \frac{\Sigma fd'}{\Sigma f} \times i = 55 + \frac{(-98)}{120} \times 10 = 55 - 8{\cdot}17$$

$$= \text{Rs. } 46{\cdot}83.$$

Example.

The following are the monthly salaries in rupees of 20 employees of a firm :—

130	125	110	100	80	62	76	98	103	122
145	151	65	71	132	118	142	116	85	95

The firm gives bonuses of Rs. 10, 15, 20, 25 and 30 for individuals in the respective salary groups : exceeding Rs. 60 but not exceeding Rs. 80 ; exceeding Rs. 80 but not exceeding Rs. 100 and so on up to exceeding Rs. 140 but not exceeding Rs. 160. Find the average bonus paid per employee.

[C.A. Nov. 1964]

From the monthly salaries of the employees, we find the number of employees lying between the salary groups mentioned as follows :

Salary (Rs.)	*No. of employees*
60— 80	4
80—100	4
100—120	5
120—140	4
140—160	3
Total	20

Calculation of average bonus

Bonus (Rs.) (x)	*No. of employees* (f)	fx
10	4	40
15	4	60
20	5	100
25	4	100
30	3	90
Total	20	390

$$\therefore \quad \text{A.M.} = \frac{\Sigma fx}{\Sigma f} = \frac{390}{20} = \text{Rs. } 19{\cdot}50.$$

Calculation of A.M. from Unequal Width of Classes.

The calculation is similar to that of the equal width of classes. The idea will be clear from the following example.

Example.

The Table given below shows the number of persons with different incomes in U. S. A. during the year 1929 :

Income in thousands of dollars	*No. of persons in lakhs*
0— 1	13
1— 2	90
2— 3	81
3— 5	117
5— 10	66
10— 25	27
25— 50	6
50— 100	2
100—1000	2

—Calculate the average income per head.

Income ('000) dollars	*No. of persons* in lakhs (*f*)	*Mid-value* (*x*)	$d=x-A$	*fd*
0— 1	13	·5	− 7	− 91
1— 2	90	1·5	− 6	−540
2— 3	81	2·5	− 5	−405
3— 5	117	4	− 3·5	−409·5
5— 10	66	7·5	0	0
10— 25	27	17·5	10	270
25— 50	6	37·5	30	180
50— 100	2	75	67·5	135
100—1000	2	550	542·5	1085
Total	404			224·5

Let A = 7·5

$$\text{A.M.} = A + \frac{\Sigma fd}{\Sigma f} = 7{\cdot}5 + \frac{224{\cdot}5}{404} = 7{\cdot}5 + {\cdot}56 = 8{\cdot}06 \text{ dollars.}$$

Calculation of A.M. in case of Open-end Classes.

Open-end classes are those in which the lower class-limit of the first class and the upper class-limit of the last class are not known. Assumption for finding such class-limits depends upon the classes following the first class up to the proceeding of the last class. For example :

Marks	*No. of students*
below 5	10
5—10	7
10—15	5
15—20	9
above 20	9

In the example, the width of classes is uniform, so the appropriate assumption would be lower limit of the first class is 0 and the upper limit of the last class is 25. Thus the first class would be 0—5 and the last class 20—25. Now mean is to be calculated as usual process.

Let us take another example :

Marks	*No. of students*
below 5	10
5—15	7
15—30	5
30—50	9
above 50	9

In the second class, width is 10, in the third class, width is 15 and in the fourth class, it is 20, *i.e.*, width is increasing by 5. So the appropriate assumption would be that the lower limit of first class is 0 and the upper limit of the last class is 75. In other words, the first class would be 0—5 and the last class 50—75.

For class-intervals of varying width, no assumption is appropriate for finding open-end class-limits. In such cases, calculation of median or mode is preferred than that of mean.

Finding of Missing Frequency.

The idea of finding the missing frequency will be clear from the following example.

The A.M. of the following frequency distribution is 1·46.

No. of accidents	0	1	2	3	4	5	Total
No. of days (frequency)	46	f_1	f_2	25	10	5	200

Find the values of f_1 and f_2.

x	f	d $=x-2$	fd
0	46	-2	-92
1	f_1	-1	$-f_1$
2	f_2	0	0
3	25	1	25
4	10	2	20
5	5	3	15
Total	200		$-32-f_1$

$$\text{A.M.}=\text{A}+\frac{\Sigma fd}{\Sigma f}$$

or, $1·46=2+\dfrac{-32-f_1}{200}$

or, $-0·54=\dfrac{-(32+f_1)}{200}$

or, $108=32+f_1$

or, $f_1=76$

Now $f_2=200-(46+76+25+10+5)$
$=38.$

Checking Accuracy of Computations.

The formula of '*Charlier's Check*' is as follows :

$$\Sigma f(d'+1)=\Sigma fd'+\Sigma f \quad [\text{or } \Sigma f(d+1)=\Sigma fd+\Sigma f]$$

If this equation is not satisfied, it means that there is some mistake in calculation.

Example.

Apply Charlier's Check in the following Table to find the mean :

Marks	0—5	5—10	10—15	15—20	20—25	Total
No. of students	10	7	5	9	9	40

Marks	*No. of students* (f)	*Mid. pt.* (x)	$d=x-A$	$d'=d/5$	fd'	$f(d'+1)$
0—5	10	2·5	−10	−2	−20	−10
5—10	7	7·5	−5	−1	−7	5
10—15	5	12·5	0	0	0	0
15—20	9	17·5	5	1	9	18
20—25	9	22·5	10	2	18	27
Total	40	—	—	—	0	40

Let A = 12·5

From, $\Sigma f(d'+1)=\Sigma fd'+\Sigma f$

Left side = 40 ; Right side = 0 + 40 = 40.

Hence the calculation is correct.

$$\text{Now, A. M.}=A+\frac{\Sigma fd'}{\Sigma f}\times i=12{\cdot}5+\frac{0}{40}\times 5=12{\cdot}5+0=12{\cdot}5 \text{ marks.}$$

Mean of Composite Group.

If $\bar{X}_1$, $\bar{X}_2$ are the means of two groups having observations N_1, N_2 respectively, then the mean ($\bar{X}$) of the composite group $N\ (=N_1+N_2)$ is given by the relation,

$$N\bar{X}=N_1\bar{X}_1+N_2\bar{X}_2.$$

Note. For three groups of respective observations N_1, N_2, N_3 and means $\overline{X}_1$, $\overline{X}_2$, $\overline{X}_3$, we have $N\overline{X} = N_1\overline{X}_1 + N_2\overline{X}_2 + N_3\overline{X}_3$.

Example.

The mean annual salary paid to all employees of a company was Rs. 5,000. The mean annual salaries paid to male and female employees were Rs. 5,200 and Rs. 4,200 respectively. Determine the percentage of males and females employed by the company.

Here, $\overline{X}_1 = 5{,}200$, $N_1 = ?$, $\overline{X}_2 = 4{,}200$, $\overline{X} = 5{,}000$, $N = N_1 + N_2$.

Now, $(N_1 + N_2)\ 5{,}000 = 5{,}200N_1 + 4{,}200N_2$

or, $5{,}000N_1 - 5{,}200N_1 = 4{,}200N_2 - 5{,}000N_2$

or, $-200N_1 = -800N_2$

or, $N_1 : N_2 = 800 : 200 = 4 : 1$

$\therefore$ Percentage of male $= \dfrac{4}{4+1} \times 100 = \dfrac{4}{5} \times 100 = 80$,

and that of female $= \dfrac{1}{4+1} \times 100 = \dfrac{1}{5} \times 100 = 20$.

Example.

Calculate the mean from the following frequency distribution :

x :	2	3	5	6—8	9—11	12—14	15—20	21—26	27—32
f :	4	1	2	3	4	3	6	3	1

The whole distribution is divided into 3 parts. For each part mean is to be calculated. Lastly to find final mean by using composite group.

Part I

x	f	fx
2	4	8
3	1	3
5	2	10
Total	7	21

$\therefore$ A.M. $= \dfrac{\Sigma fx}{\Sigma f} = \dfrac{21}{7} = 3$

Part II

Group	f	x	fx
6— 8	3	7	21
9—11	4	10	40
12—14	3	13	39
Total	10	—	100

$\therefore$ A.M. $= \dfrac{100}{10} = 10$.

Part III

Group	*f*	*x*	*fx*
15—20	6	17·5	105·0
21—26	3	23·5	70·5
27—32	1	29·5	29·5
Total	10	—	205·0

$\therefore$ A.M. $= \frac{205}{10} = 20{\cdot}5$

Now, arranging the data for 3 parts, we find

$N_1 = 7$ $\quad N_2 = 10$ $\quad N_3 = 10$ $\quad N = 7 + 10 + 10 = 27$

$\overline{X}_1 = 3$ $\quad \overline{X}_2 = 10$ $\quad \overline{X}_3 = 20{\cdot}5$ $\quad \overline{X} = ?$

We find, $27X = 7 \times 3 + 10 \times 10 + 10 \times 20{\cdot}5$

$= 21 + 100 + 205 = 326$

$$\therefore \quad \overline{X} = \frac{326}{27} = 12{\cdot}07$$

Advantages and Disadvantages of Arithmetic Mean.

Advantages :

(i) It is easy to calculate and simple to understand.

(ii) For counting mean, all the data are utilised. It can be determined even when only the number of items and their aggregate are known.

(iii) It is capable of further mathematical treatment.

(iv) It provides a good basis to compare two or more frequency distribution.

(v) Mean does not necessitate the arrangement of data.

Disadvantages :

(i) It may give considerable weight to extreme items. Mean of 2, 6, 301 is 103 and none of the value is adequately represented by the mean 103.

(ii) In some cases, arithmetic mean may give misleading impressions. For example, average number of patients admitted in a hospital is 10·7 per day. Here mean is a useful information, but does not represent the actual item.

(iii) It can hardly be located by inspection.

Geometric Mean (G.M.).

Definition : The geometric mean (G) of the n positive values of a variate $x_1\ x_2,\ x_3,\ \ldots,\ x_n$ is the n root of the product of the values, *i.e.*,

$G=\sqrt[n]{x_1 .\ x_2 .\ \ldots\ x_n}$. It means,

$G=(x_1 \cdot x_2 \cdot \cdots x_n)^{\frac{1}{n}}$. Now taking logarithms on both sides we find,

$$\log G=\frac{1}{n}\log (x_1 .\ x_2 .\ \cdots .\ x_n)=\frac{1}{n}(\log x_1+\log x_2+\cdots+\log x_n)$$

$$=\frac{1}{n}\Sigma \log x \ldots(1) \qquad \therefore\ G=\text{antilog}\left[\frac{1}{n}\Sigma \log x\right]$$

Thus, from formula (1) we find that the logarithm of the G.M. of $x_1, x_2, \cdots, x_n=$ A.M. of logarithms of $x_1, x_2, \cdots, x_n$.

Properties.

1. The product of n values of a variate is equal to the n-th power of their G.M.

i.e., $x_1 . x_2 . \cdots x_n=G^n$ (it is clear from the definition).

2. The logarithm of G.M. of n observations is equal to the A.M. of logarithms of n observations. [Formula (1) states it].

3. The product of the ratios of each of the n observations to the G.M. is always unity.

Taking G as geometric mean of n observations $x_1,\ x_2,\ \ldots,\ x_n$ the ratios of each observation to the geometric mean are

$$\frac{x_1}{G}, \frac{x_2}{G}, \ldots, \frac{x_n}{G}.$$

By definition, $G=\sqrt[n]{x_1 .\ x_2 .\ \ldots\ x_n}$ or $G^n=(x_1 .\ x_2 .\ \ldots\ x_n)$. Now the product of the ratios,

$$\frac{x_1}{G} . \frac{x_2}{G} . \ldots \frac{x_n}{G}=\frac{x_1 .\ x_2 .\ \ldots\ x_n}{G.G.\ \ldots \text{ to } n \text{ times}}=\frac{G^n}{G^n}=1.$$

4. If $G_1, G_2, \ldots$ are the geometric means of different groups having observations $n_1,\ n_2,\ \ldots$ respectively, then G.M. (G) of composite group is given by

$$G=\sqrt[N]{G_1^{n_1} .\ G_2^{n_2} .\ \cdots}$$

where $N=n_1+n_2+\cdots$

i.e., $\log G=\frac{1}{N}\left[n_1 \log G_1+n_2 \log G_2+\cdots\cdots\right]$.

Example.

Find the G. M. of 111, 171, 191, 212.

If G indicates the G. M. of the numbers, then

$$G=\sqrt[4]{111\times171\times191\times212} \qquad \text{here } n=4$$

$$\begin{aligned}\log\ G&=\tfrac{1}{4}(\log 111+\log 171+\log 191+\log 212)\\ &=\tfrac{1}{4}(2·0453+2·2330+2·2810+2·3263)\\ &=\tfrac{1}{4}(8·8856)=2·2214\end{aligned}$$

$\therefore\ G=$ antilog $2·2214=166·5.$

Weighted Geometric Mean.

Computation of geometric mean of numbers when they are weighted respectively, will be cleared from the following example.

Example.

Find the G. M. of 111, 171, 191, 212 having weighted by 3, 2, 4, 5 respectively.

x		$\log x$	$f \log x$
111	3	2·0453	6·1359
171	2	2·2330	4·4660
191	4	2·2810	9·1240
212	5	2·3263	11·6315
Total	14	......	31·3574

$$\log\ G=\frac{\Sigma f \log x}{\Sigma f}=\frac{31·3574}{14}=2·2391$$

$\therefore\ G=$ antilog $2·2391=173·4.$

Example.

The weighted geometric mean of the four numbers 8, 25, 17 and 30 is 15·3. If the weights of the first three numbers are 5, 3 and 4 respectively, find the weight of the fourth number.

[I.C.W.A. Jan. 1971]

Taking f_4 as the weight of the fourth number 30, we find,

x	f	$\log x$	$f \log x$
8	5	0·9031	4·5155
25	3	1·3979	4·1937
17	4	1·2304	4·9216
30	f_4	1·4771	1·4771 f_4
	$12+f_4$		$13{\cdot}6308+1{\cdot}4771\,f_4$

Now, $\log G = \dfrac{\Sigma f \log x}{\Sigma f}$

or, $\log 15{\cdot}3 = \dfrac{13{\cdot}6308+1{\cdot}4771\,f_4}{12+f_4}$

or, $1{\cdot}1847 = \dfrac{13{\cdot}6308+1{\cdot}4771\,f_4}{12+f_4}$

or, $(1{\cdot}1847)(12+f_4) = 13{\cdot}6308+1{\cdot}4771\,f_4$

or, $14{\cdot}2164+1{\cdot}1847\,f_4 = 13{\cdot}6308+1{\cdot}4771\,f_4$

or, $14{\cdot}2164-13{\cdot}6308 = 1{\cdot}4771\,f_4 - 1{\cdot}1847 f_4$

or, $\cdot 5856 = \cdot 2924\,f_4 \quad \therefore \quad f_4 = \dfrac{\cdot 5856}{\cdot 2924} = 2.$

Advantages and Disadvantages of Geometric Mean.

Advantages :

(i) It is not influenced by the extreme items to the same extent as mean.

(ii) It is rigidly defined and its value is a precise figure.

(iii) It is based on all observations and capable of further algebraic treatment.

(iv) It is useful in calculating index numbers.

Disadvantages :

(i) It is neither easy to calculate nor it is simple to understand.

(ii) If any value of a set of observations is zero, the geometric mean would be zero, and it cannot be determined.

(iii) If again any value becomes negative, geometric mean becomes imaginary.

Uses of Geometric Mean.

1. It is used to find average of the rates of changes. If the prices, for the years 1970 to 1972, be increased by 5%, 7%, and 12% respectively, then average annual increase is not 8% $\left(=\frac{5+7+12}{3}=8\right)$ as determined by mean, but 7·5% (G.M. of 5, 7, 12) as determined by geometric mean. Geometric mean is useful in measuring growth of population, as population increases in geometric progression.

2. It is considered to be the best average for the construction of index numbers.

Harmonic Mean (H.M.).

Definition : The Harmonic Mean (H.) for n observations x_1, x_2, $\cdots$, x_n is the total number divided by the sum of the reciprocals of the numbers.

$$i.e., \quad \text{H.} = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}} = \frac{n}{\Sigma\frac{1}{x}}$$

Again, $\frac{1}{\text{H}} = \frac{\Sigma\frac{1}{x}}{n}$ (*i.e.*, reciprocal of H.M. = A.M. of reciprocals of the numbers).

Example.

Find the H.M. of 3, 6, 12 and 15.

$$\text{H.M.} = \frac{4}{\frac{1}{3}+\frac{1}{6}+\frac{1}{12}+\frac{1}{15}} = \frac{4}{\frac{20+10+5+4}{60}}$$

$$= \frac{4}{\frac{39}{60}} = \frac{4\times60}{39} = \frac{240}{39} = 6\frac{6}{39}.$$

Example.

Find the H.M. of $1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n}$

$$\text{H.M.} = \frac{n}{1+2+3+\cdots+n} = \frac{n}{\frac{n}{2}(2+n-1)} = \frac{2n}{n(n+1)} = \frac{2}{n+1}.$$

Note : The denominator is in A.P., use $S=\frac{n}{2}\{2a+(n-1)d\}$.

Example.

A motor car covered a distance of 50 miles four times. The first time at 50 m.p.h., the second at 20 m.p.h., the third at 40 m.p.h., and the fourth at 25 m.p.h. Calculate the average speed and explain the choice of the average. [C.A. Nov. 1967]

$$\text{Average Speed (H.M.)} = \frac{4}{\frac{1}{50}+\frac{1}{20}+\frac{1}{40}+\frac{1}{25}} = \frac{4}{\frac{20+50+25+40}{1000}}$$

$$= 4 \times \frac{1000}{135} = \frac{800}{27} = 29{\cdot}63 = 30 \text{ (app.) m.p.h.}$$

For the statement x *units per hour*, when the different values of x (*i.e.*, *distances*) are given, to find average, use H.M. If again hours (*i.e.*, *time of journey*) are given, to find average, we are to use A.M. In the above example, miles (*distances*) are given, so we have used H.M.

Weighted H.M. The formula to be used is as follows :

$$\text{H.M.} = \frac{N}{\frac{f_1}{x_1}+\frac{f_2}{x_2}+\cdots+\frac{f_n}{x_n}}, \quad \text{where} \quad \Sigma f = N.$$

Example.

(a) A person travelled 20 k.m. at 5 k.m.p.h. and again 24 k.m. at 4 k.m.p.h., to find average speed.

(b) A person travelled 20 hours at 5 k.m.p.h. and again 24 hours at 4 k.m.p.h., to find average speed.

(a) We are to apply H.M. (weighted) in this case, since distances are given.

$$\begin{array}{c}\text{Average speed}\\ \text{(H.M.)}\end{array} = \frac{20+24}{\frac{20}{5}+\frac{24}{4}} = \frac{44}{4+6} = \frac{44}{10} = 4{\cdot}4 \text{ k.m.p.h.}$$

(b) We are to apply A.M. (weighted), since times of journey are given.

$$\begin{array}{c}\text{Average speed}\\ \text{(A.M.)}\end{array} = \frac{20\times5+24\times4}{20+24} = \frac{100+96}{44} = \frac{196}{44} = 4{\cdot}45 \text{ (app.). k.m.p.h.}$$

Advantages and Disadvantages of Harmonic Mean.

Advantages :

(i) Like A.M. and G.M. it is also based on all observations.

(ii) Capable of further algebraic treatment.

(iii) It is extremely useful while averaging certain types of rates and ratios.

Disadvantages :

(i) It is not readily understood nor can it be calculated with ease.

(ii) It is usually a value which may not be a member of the given set of numbers.

(iii) It cannot be calculated when there are both negative and positive values in a series or, one or more values is zero.

Relations between A.M., G.M. and H.M.

(1) The Arithmetic Mean is never less than the Geometric Mean, again Geometric Mean is never less than the Harmonic Mean.

[I. C.W.A. June '79]

i.e., $$\text{A.M.} \geqslant \text{G.M.} \geqslant \text{H.M.}$$

For the observations x_1 and x_2, we know

$$(\sqrt{x_1} - \sqrt{x_2})^2 \geqslant 0 \quad \text{or,} \quad x_1 + x_2 - 2\sqrt{x_1 x_2} \geqslant 0$$

$$\text{or,} \quad x_1 + x_2 \geqslant 2\sqrt{x_1 x_2} \quad \text{or,} \quad \frac{x_1 + x_2}{2} \geqslant \sqrt{x_1 x_2} \text{ or, A.M.} \geqslant \text{G.M.}$$

This is for two observations only. Similarly for the other observations x_3, x_4 we can show $\frac{x_3 + x_4}{2} \geqslant \sqrt{x_3 x_4}$. Again for the observations $\frac{x_1 + x_2}{2}$ and $\frac{x_3 + x_4}{2}$ we can show (similarly)

$$\frac{x_1 + x_2 + x_3 + x_4}{4} \geqslant \sqrt{\frac{x_1 + x_2}{2} \cdot \frac{x_3 + x_4}{2}}.$$

$$\text{As} \quad \frac{x_1 + x_2}{2} \cdot \frac{x_3 + x_4}{2} \geqslant \sqrt{x_1 . x_2}\, \sqrt{x_3 . x_4}.$$

$$\therefore \quad \frac{x_1 + x_2 + x_3 + x_4}{4} \geqslant \sqrt{\sqrt{x_1 x_2} \cdot \sqrt{x_3 x_4}}$$

$$\text{or} \quad \frac{x_1 + x_2 + x_3 + x_4}{4} \geqslant \sqrt[4]{x_1 x_2 x_3 x_4}$$

$\therefore$ A.M. $\geqslant$ G.M. (this is for four observations).

In this way, for any number of observation, we have A.M. $\geqslant$ G.M.

Again for $\frac{1}{x_1}$ and $\frac{1}{x_2}$ (observations)

$$\left(\sqrt{\frac{1}{x_1}} \sim \sqrt{\frac{1}{x_2}}\right)^2 \geqslant 0 \quad \text{or,} \quad \frac{1}{x_1} + \frac{1}{x_2} - \frac{2}{\sqrt{x_1 x_2}} \geqslant 0$$

or, $\frac{1}{x_1}+\frac{1}{x_2} \geqslant \frac{2}{\sqrt{x_1x_2}}$ or, $\sqrt{x_1x_2} \geqslant \frac{2}{\frac{1}{x_1}+\frac{1}{x_2}}$

or, G.M. $\geqslant$ H.M. (This is true for any number of observations)

$\therefore$ A.M. $\geqslant$ G.M. $\geqslant$ H.M.

(2) For a pair of observations only, $\frac{\text{A.M.}}{\text{G.M.}}=\frac{\text{G.M.}}{\text{H.M.}}$

or, $(\text{G.M.})^2=\text{A.M.}\times\text{H.M.}$

i.e., the Geometric Mean for a pair of observations is the geometric mean of their Arithmetic and Harmonic Means.

Median.

Definition : If a set of observations are arranged in order of magnitude (ascending or descending), then the middle most or central value gives the median.

Median divides the observations in two equal parts, in such a way that the number of observations smaller than median is equal to the number greater than it. It is not effected by extremely large or small observations. Median is, thus, an average of position. In certain sense, it is the real measure of central tendency.

Calculation of Median.

(A) ***For Series of Individual Observations**** :

At first, the given data are to be arranged in order of magnitude (ascending or descending). Now for n (the total numer of items) odd,

median = value of $\frac{n+1}{2}$th item and for n even,

median = average value of $\frac{n}{2}$ th item and $\frac{n+1}{2}$ th item.

(*i.e.*, the next item)

Note. $\frac{n+1}{2}$ th item gives the location of median, but not its magnitude.

Example.

To find the median of the following marks obtained by 7 students :

4, 12, 7, 9, 14, 17, 16

* Individual observations are those observations (or variates) having no frequencies or frequency is unit in every case.

(i) Arrangement of marks : 4, 7, 9, 12, 14, 16, 17

(ii) $n = 7 =$ an odd number

(iii) Median = value of $\frac{n+1}{2}$th item

= „ „ $\frac{7+1}{2}$th „

= „ „ 4th „

= 12 (from the arranged data)

∴ median mark is 12.

Example.

To find the median of marks :

4, 12, 7, 9, 14, 17, 16, 21.

(i) Arrangement : 4, 7, 9, 12, 14, 16, 17, 21.

(ii) $n = 8 =$ an even number.

(iii) Median = average value of $\frac{n}{2}$ th item and the next item

= „ „ „ $\frac{8}{2}$ th „ „ „ „ „

= „ „ „ 4th „ „ „ 5th „

= average value of 12 and 14 marks $= \frac{12+14}{2}$

= 13 marks.

(B) ***For Discrete Series*** (*Simple Frequency Distribution*) :

Cumulative frequency (less than type) is calculated. Now the value of the variable corresponding to the cumulative frequency $\frac{N+1}{2}$ gives the median, when N is the total frequency.

Example.

To find the median of the following frequency distribution :

x :	1	2	3	4	5	6
f :	7	12	17	19	21	24

x	f	*Cum. freq.*
1	7	7
2	12	19
3	17	36
4	19	55
5	21	76
6	24	100 (= N)

Now, median = value of $\frac{N+1}{2}$th item,

= " " $\frac{100+1}{2}$th "

= " " 50·5th "

From the last column, it is found 50·5 is greater than the cumulative frequency 36, but less than the next cum. freq. 55 corresponding to $x=4$. All the 19 items (from 37 to 55) have the same variate 4. And 50·5th item is also one of these 19 item. ∴ Median = 4.

(C) ***For Continuous Series*** (*Grouped Frequency Distribution*) :

We are to determine the particular class in which the value of the median lies, by using the formula $\frac{N}{2}$ (and not by $\frac{N+1}{2}$, as in continuous series $\frac{N}{2}$ divides the area of the curve into two equal parts). After locating median, its magnitude is measured by applying the formula of interpolation given below—

$$\text{Median} = l_1 + \frac{l_2 - l_1}{f}\left(m - c\right), \text{ where } m = \frac{N}{2}.$$

Where, l_1 = lower limit of the class in which median lies,

l_2 = upper limit of the class in which median lies,

f = the frequency of the class in which median falls,

m = middle item (*i.e.*, item at which median is located or $\frac{N}{2}$th item),

c = cumulative frequency of the class preceding the median class.

Note. The above formula is based on the assumption that the frequencies of the class-interval in which median lies are uniformly distributed over the entire class-interval.

Example.

Find the median and median-class of the data given below :

Class-boundaries	15—25	25—35	35—45	45—55	55—65	65—75
Frequency	4	11	19	14	0	2

[I. C. W. A. Jan. 1965]

Class-boundaries	*Frequency*	*Cumulative frequency*
15—25	4	4
25—35	11	15
35—45	19	34
45—55	14	48
55—65	0	48
65—75	2	50 (= N)

Median = value of $\frac{N}{2}$th item = value of $\frac{50}{2}$th item

= ,, ,, 25th item, which is greater than cum. freq. 15 and less than cum. freq. 34. So median lies in the class 35—45.

Now, median $= l_1 + \frac{l_2 - l_1}{f}(m - c)$, where $l_1 = 35,\ l_2 = 45,\ f = 19$
$m = 25,\ c = 15$

$$= 35 + \frac{45 - 35}{19}(25 - 15)$$

$$= 35 + \frac{10}{19} \times 10 = 35 + \frac{100}{19} = 35 + 5{\cdot}26 = 40{\cdot}26$$

∴ reqd. median is 40·26 and median-class is (35 – 45).

Example.

Compute the median from the following data :

Mid-value	*Frequency*	*Mid-value*	*Frequency*
115	6	165	60
125	25	175	38
135	48	185	22
145	72	195	3
155	116		

At first we are to find the class-boundaries from the mid-values given.

Class-boundaries	*Frequency*	*Cumulative frequency*
110—120	6	6
120—130	25	31
130—140	48	79
140—150	72	151
150—160	116	267
160—170	60	327
170—180	38	365
180—190	22	387
190—200	3	390 (=90)

Median = value of $\frac{N}{2}$th item = value of $\frac{390}{2}$th item

= value of 195th item, so median lies in the class (150—160).

Again, median $= l_1 + \frac{l_2 - l_1}{f}(m - c)$,

$$l_1 = 150,\ l_2 = 160,\ f = 116,\ m = 195,\ c = 151$$

$$= 150 + \frac{160 - 150}{116}(195 - 151)$$

$$= 150 + \frac{10}{116} \times 44 = 150 + 3{\cdot}79 = 153{\cdot}79 \doteq 153{\cdot}8 \quad \text{(app.)}$$

Example.

The following is the Table which gives you the distribution of marks secured by some students in an examination :

Marks :	0—20	21—30	31—40	41—50	51—60	61—70	71—80
No. of students :	42	38	120	84	48	36	31

Find : (i) median marks, (ii) the percentage of failure if the minimum for a pass is 35 marks.

[C. A. Nov. 1969]

We are to make the class-boundaries from the class-limits given and then to find the cumulative frequency.

Class-boundaries	*Frequency*	*Cumulative frequency*
0—20·5	42	42
20·5—30·5	38	80
30·5—40·5	120	200
40·5—50·5	84	284
50·5—60·5	48	332
60·5—70·5	36	368
70·5—80·5	31	399 (= N)

Median = value of $\frac{N}{2}$th item

= " " 199·5 "

Median lies in (30·5 – 40·5)

Now, $l_1 = 30{\cdot}5$, $l_2 = 40{\cdot}5$,

$f = 120$, $m = 199{\cdot}5$, $c = 80$

$$\text{Median} = l_1 + \frac{l_2 - l_1}{f}(m - c)$$

$$= 30{\cdot}5 + \frac{40{\cdot}5 - 30{\cdot}5}{120}(199{\cdot}5 - 80)$$

$$= 30{\cdot}5 + \frac{10}{120} \times 119{\cdot}5$$

$$= 30{\cdot}5 + 9{\cdot}96 = 40{\cdot}46.$$

The mark 35 represents the interval (34·5 – 35·5) taking marks as a continuous variable. Minimum pass mark is 34·5. Number of students (F) obtaining less than 34·5 is the cumulative frequency corresponding to 34·5 marks. Now using the above median formula, putting respective values, we find,

$$34{\cdot}5 = 30{\cdot}5 + \frac{40{\cdot}5 - 30{\cdot}5}{120}(F - 80)$$

or, $$34{\cdot}5 = 30{\cdot}5 + \frac{10}{120}(F - 80)$$

or, $$34{\cdot}5 - 30{\cdot}5 = \frac{1}{12}(F - 80)$$

or, $$4 = \frac{F - 80}{12}$$

or, $$F - 80 = 48$$

$$\therefore \quad F = 128$$

$\therefore$ reqd. percentage of failure $= \frac{128}{399} \times 100 = 32{\cdot}08\%.$

Note. In the first class, 0 is taken as lower boundary as there cannot be any number less than zero.

Calculation of Median when Class-intervals are unequal.

In such cases, the frequencies need not be adjusted to make the class-intervals equal. Formula of median is to be applied directly.

Example.

To calculate median from the following Table :

Marks :	0—10	10—30	30—60	60—70	70—90
No. of students :	15	25	30	4	10

Marks	*f*	*Cum. freq.*
0—10	15	15
10—30	25	40
30—60	30	70
60—70	4	74
70—90	10	84 (= N)

Median = value of $\frac{84}{2}$th item

= „ „ 42th „

So, median lies in the class (30—60)

Here, $l_1 = 30$, $l_2 = 60$, $f = 30$, $m = 42$, $c = 40$

$$\therefore \quad \text{Median} = 30 + \frac{60-30}{30} \times (42-40)$$

$$= 30 + \frac{30}{30} \times 2 = 30 + 2$$

= 32 marks.

We will get the same result if the class-intervals are made equal :

Marks	*f*	*Cum. freq.*
0—10	15	15
10—20	12·5	27·5
20—30	12·5	40
30—40	10	50
40—50	10	60
50—60	10	70
60—70	4	74
70—80	5	79
80—90	5	84 (= N)

Median lies in the class (30—40)

$l_1 = 30$, $l_2 = 40$, $f = 10$, $m = 42$, $c = 40$

$$\text{Median} = 30 + \frac{40-30}{10} (42-40)$$

$$= 30 + \frac{10}{10} \times 2 = 30 + 2$$

= 32 marks.

(D) *Graphic Method.*

Median can be determined graphically by the following methods :

(i) Draw *less than* (or *greater than*) type ogive, taking the variation X-axis and the cumulative frequency on Y-axis. Now corresponding to N/2 on Y-axis draw a horizontal line to meet at ogive, and again from the point of intersection, perpendicular is now drawn on X-axis. The point on X-axis is read off, which gives the median.

Example. To find the median graphically from the following Table :

Wages (Rs.) :	10—20	20—30	30—40	40—50	50—60	60—70	70—80
No. of workers :	5	10	12	16	8	5	4

Wages (Rs.) (less than)	*No. of workers*	*Wages* (Rs.) (greater than)	*No. of workers*
20	5	10	60
30	15	20	55
40	27	30	45
50	43	40	33
60	51	50	17
70	56	60	9
80	60	70	4
		80	0

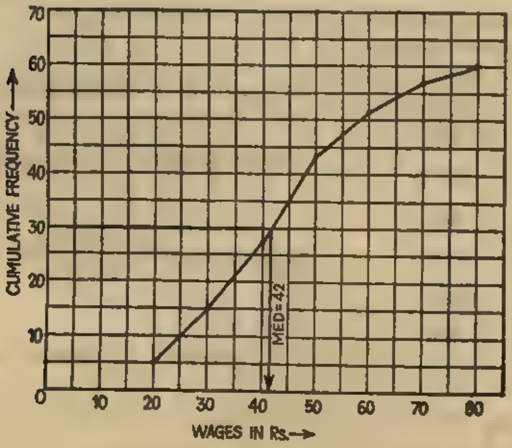


Fig. 31

We draw less than type ogive, as shown before.

Median = size of $\frac{N}{2}$th item = size of 30th item. Now take 30 on Y-axis, and from 30 draw a horizontal line to meet the ogive. From this point on ogive, draw a perpendicular on the X-axis. The point on X-axis is read off. The point is 42, which gives the median. So median is Rs. 42.

Note. If we draw greater than type ogive, we would get the same result.

(ii) Draw two ogives. From the point of intersection of the curves (*i.e.*, ogives), draw a perpendicular to meet the X-axis. The point on the X-axis is read off, which gives the median.

Example.

The data given in the above example are taken here into consideration :

Draw two ogives (less than type and greater than type). From the point of intersection, draw a perpendicular to the X-axis. The point on X-axis shows 42. So median is Rs. 42.

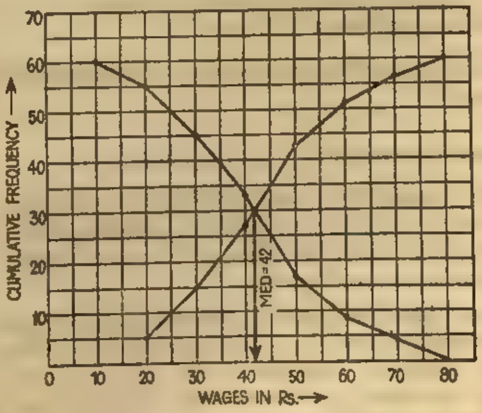


Fig. 32

Advantages and Disadvantages of Median.

Advantages :

(i) The median, unlike the mean, is uneffected by the extreme values of the variable.

(ii) It is easy to calculate and simple to understand, particularly in a series of individual observations and a discrete series.

(iii) It is capable of further algebraic treatment. It is used in calculating mean deviation.

(iv) It can be located by inspection, after arranging the data in order of magnitude.

(v) Median can be calculated even if the items at the extreme are not known, but if we know the central items and the total number of items.

(vi) It can be determined graphically.

Disadvantages :

(i) For calculation, it is necessary to arrange the data, other averages do not need any such arrangement.

(ii) It is amenable to algebraic treatment in a limited sense. Median cannot be used to calculate the combined median of two or more groups, like mean.

(iii) It cannot be computed precisely when it lies between two items.

(iv) Process involved to calculate median in case of continuous series is difficult to follow.

(v) Median is effected more by sampling fluctuations than the mean.

Other Measures (*Regarding the median principle*).

It has been seen that median divides an arrayed series in two equal parts. Now for further study of composition of a series, it may be divided into four, five, six, ten or hundred parts. Usually it is divided four, ten or hundred parts.

Just as we have one median dividing a series in two equal parts, so three items would divide it in four parts, nine items in ten parts and ninety-nine items in hundred parts. The values of these items are respectively known as *Quartiles*, *Deciles* and *Percentiles*. Quintiles, Septiles and Octiles divide a series respectively in five, seven and eight parts.

Thus we find three quartiles, nine deciles and ninety-nine percentiles in a series. The second quartiles, fifth decile and fifty-th percentile is median. *First Quartile* or *Lower Quartile* (Q_1) is that value of the variable which divides the first half of a series in two equal parts. *Third Quartile* or *Upper Quartile* (Q_3) is the value of the variable that divides the latter half of a series.

The calculation of Quartiles, Deciles and Percentiles and other such values is done by the same rules applied in calculating the median.

(A) *For Series of Individual Observation.*

The data are to be arranged in increasing order of magnitude :

1st Quartile, $Q_1 =$ size of $\frac{n+1}{4}$th item

3rd Quartile, $Q_3 =$ " " $\frac{3(n+1)}{4}$th item

1st Decile, $D_1 =$ " " $\frac{n+1}{10}$th item

7th Decile, $D_7 =$ " " $\frac{7(n+1)}{10}$th item

K-th Decile, $D_k =$ " " $\frac{k(n+1)}{10}$th item (for $k = 1, 2, \ldots\ldots, 8, 9$)

1st Percentile $P_1 =$ " " $\frac{n+1}{100}$th item

K-th Percentile $P_k =$ " " $\frac{k(n+1)}{100}$th item (for $k = 1, 2, \ldots\ldots, 98, 99$)

Example.

To find Q_1, Q_3, D_4 and P_{60} from the following weights (in Kg.) :

19,	27,	24,	39,	57,	44,	56,	50,	59,	67,
62,	42,	47,	60,	26,	34,	57,	51,	59,	45.

Arrangement :

Serial no.	*Weight* (Kg.)	*Serial no.*	*Weight* (Kg.	*Serial no.*	*Weight* (Kg.)
1	19	8	44	15	57
2	24	9	45	16	59
3	26	10	47	17	59
4	27	11	50	18	60
5	34	12	51	19	62
6	39	13	56	20	67
7	42	14	57		

Here $n = 20$.

Q_1 (first quartile) = size of $\frac{n+1}{4}$th item = size of $\frac{20+1}{4}$th item

= size of 5·25th item

= size of 5th item + $\frac{1}{4}$ (size of 6th item − size of 5th item)

$= 34 + \frac{1}{4}(39 - 34) = 34 + 1{\cdot}25 = 35{\cdot}25$ Kg.

Q_3 (third quartile) = size of $\frac{3(n+1)}{4}$th item = size of $\frac{3(20+1)}{4}$th item

= size of 15·75th item

= size of 15th item + $\frac{3}{4}$ (size of 16th item − size of 15th item)

$= 57 + \frac{3}{4}(59 - 57) = 57 + 1{\cdot}50 = 58{\cdot}50$ Kg.

D_4 (fourth decile) = size of $\frac{4(n+1)}{10}$th item = size of $\frac{4(20+1)}{10}$th item

= size of 8·4th item

= size of 8th item + $\frac{4}{10}$ (size of 9th item − size of 8th item)

$= 44 + \frac{4}{10}(45 - 44) = 44 + {\cdot}4 = 44{\cdot}4$ Kg.

P_{60} (sixty-th percentile) = size of $\frac{60(n+1)}{100}$th item

= size of $\frac{60(20+1)}{100}$th item

= size of 12·6th item

= size of 12th item + $\frac{6}{10}$ (size of 13th item − size of 12th item)

$= 51 + \frac{6}{10}(56 - 51) = 51 + 3 = 54$ Kg.

(B) ***For Discrete Series.***

Q_1 = size of $\frac{N+1}{4}$th item (where N is the total frequency)

= „ „ $\frac{98+1}{4}$th item (from the Table given after)

= „ „ 24·75th „ = 50 Kg.

Q_3 = size of $\frac{3(N+1)}{4}$th item

= „ „ $\frac{3(98+1)}{4}$th item (from the Table given)

= „ „ 74·25th item = 60 Kg.

D_4 = size of $\frac{4(N+1)}{10}$th item

= „ „ $\frac{4(98+1)}{10}$th item (from the Table given below)

= „ „ 39·6th item = 54 Kg.

P_{60} = size of $\frac{60(N+1)}{100}$th item

= „ „ $\frac{60(98+1)}{100}$th item (from the Table given below)

= „ „ 59·4th item = 59 Kg.

Weight (Kg.)	*Frequency*	*Cumulative frequency*
40	2	2
42	6	8
45	8	16
50	10	26
51	6	32
54	14	46
56	12	58
59	8	66
60	14	80
62	12	92
64	6	98 (= N)

(C) *For Continuous Series.*

Like median, the values of quartiles, deciles and percentiles lie in various class-intervals and the actual values are to be calculated by applying interpolation formulae.

Q_1 = size of $\frac{N}{4}$th item

= „ „ $\frac{84}{4}$th „ (from the Table given at page 151)

= „ „ 21st „, which lies in the class (36—40)

Now, $l_1=36$, $l_2=40$, $f=8$,

q (item in which quartile is located) $=21$, $c=20$.

$$\therefore \quad Q_1=l_1+\frac{l_2-l_1}{f}(q-c)=36+\frac{40-36}{8}(21-20)$$

$$=36+\frac{4}{8}=36+{}^{\cdot}5=36{}^{\cdot}5 \text{ Kg.}$$

$$Q_3=\text{size of } \frac{3N}{4}\text{th item}$$

$$= \text{ ,, ,, } \frac{3\times 84}{4}\text{th item (from the Table given after)}$$

$$= \text{ ,, ,, 63rd ,, So, } Q_3 \text{ lies in the class (52—56).}$$

Now, $l_1=52$, $l_2=56$, $f=10$, $q=63$, $c=62$

$$\therefore \quad Q_3=l_1+\frac{l_2-l_1}{f}(q-c)=52+\frac{56-52}{10}(63-62)$$

$$=52+{}^{\cdot}4=52{}^{\cdot}4 \text{ Kg.}$$

$$D_4=\text{size of } \frac{4N}{10}\text{th item}=\text{size of } \frac{4\times 84}{10}\text{th item}$$

(from the Table given)

$$= \text{ ,, ,, } 33{}^{\cdot}6\text{th ,, So } D_4 \text{ lies in the class (40—44).}$$

$$\therefore \quad D_4=40+\frac{44-40}{6}(33{}^{\cdot}6-28)=40+\frac{4}{6}\times 5{}^{\cdot}6=40+3{}^{\cdot}7$$

$$=43{}^{\cdot}7 \text{ Kg.}$$

$$P_{60}=\text{size of } \frac{60N}{100}\text{th item}$$

$$= \text{ ,, ,, } \frac{60\times 84}{100}\text{th ,, (from the Table given)}$$

$$= \text{ ,, ,, } 50{}^{\cdot}4\text{th ,, } P_{60} \text{ lies in the class (48—52).}$$

$$\therefore \quad P_{60}=48+\frac{52-48}{12}(50{}^{\cdot}4-50)=48+\frac{4}{12}\times({}^{\cdot}4)$$

$$=48+1{}^{\cdot}3=49{}^{\cdot}3 \text{ Kg.}$$

Weight (Kg.)	*Frequency*	*Cumulative frequency*
20—24	2	2
24—28	3	5
28—32	5	10
32—36	10	20
36—40	8	28
40—44	6	34
44—48	16	50
48—52	12	62
52—56	10	72
56—60	7	79
60—64	5	84

(D) *By Graphic Method.*

Like median, quartiles, deciles and percentiles can also be calculated graphically with the help of cumulative frequency curves, known as ogives. The process of estimating quartiles is shown below :

Example.

The following are the marks obtained by 123 students in statistics :

Marks obtained	*No. of students*
1— 5	7
6—10	10
11—15	16
16—20	32
21—25	24
26—30	18
31—35	10
36—40	5
41—45	1
Total	123

—Draw an ogive and locate the first and third quartiles.

At first we make the class-boundaries and cumulative frequency as shown below :

Class-boundaries	*Frequency*	*Cumulative frequency* (less than type)
'5— 5'5	7	7
5'5—10'5	10	17
10'5—15'5	16	33
15'5—20'5	32	65
20'5—25'5	24	89
25'5—30'5	18	107
30'5—35'5	10	117
35'5—40'5	5	122
40'5—45'5	1	123

Drawing of ogive (less than type) and estimation of Q_1 and Q_3 are shown below :

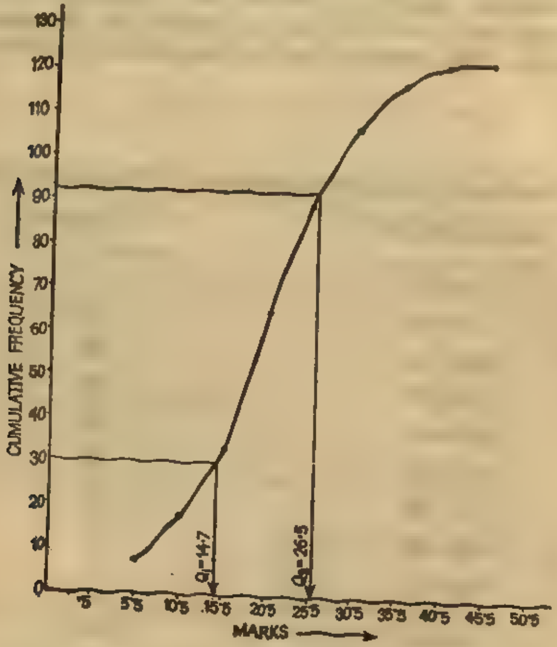


Fig. 33

We find from the above drawing,

$$Q_1 = 14{\cdot}7 \text{ marks} ;\ Q_3 = 26{\cdot}5 \text{ marks.}$$

Mode.

Definition : Mode is the value of the variate which occurs most frequently. It represents the most frequent value of a series.

When one speaks of the 'average wage', 'average student', etc., we generally mean the modal wage, the modal student. If we say that the modal wages obtained by workers in a factory are Rs. 70, we mean that the largest number of workers get the same amount. As high as Rs. 100 and as low as Rs. 50 as wages are much less frequented and they are non-modal.

Calculation : Mode cannot be determined from a series of individual observations unless it is converted to a discrete series (or continuous series). In a discrete series the value of the variate having the maximum frequency is the mode. In continuous series, the class-interval having the maximum frequency is the modal class. However the exact location of mode is done by interpolation formula like median.

Location of modal value in case of discrete series is possible if there is concentration of items at one point. If again there are two or more values having same maximum frequencies (*i.e.*, more concentrations), it becomes difficult to determine mode. Such items are known as *bi-modal*, *tri-modal* or *multi-modal* according as the items concentrate at 2, 3 or more values.

(A) ***For Individual Observations.***

The individual observations are to be first converted to discrete series (if possible). Then the variate having the maximum will be the mode.

Example. Calculate mode from the data (given) :

(Marks) : 10, 14, 24, 27, 24, 12, 11, 17.

Marks	*Frequency*
10	1
11	1
12	1
14	1
17	1
24	2
27	1

(Individual observations are converted into a discrete series)

Here marks 24 occurs maximum number of times, *i.e.*, 2. Hence the modal marks are 24, or, mode = 24 marks.

Alternatively :

Arranging the numbers : 10, 11, 12, 14, 17, (24, 24), 27.

Now 24 occurs maximum number, *i.e.*, 2.

∴ mode = 24 marks.

Note. When there are two or more values having the same maximum frequency, then mode is ill-defined. Such a sense is known as bi-modal or multi-modal as the case may be.

Example.

Marks obtained : 24, 14, 20, 17, 20, 14.

Marks	*Frequency*
14	2
17	1
20	2
24	1

Here 14 occurs 2 times (max.) and 20 occurs 2 times (max.)
∴ mode is ill-defined.

(B) ***For Discrete Series.***

To find the mode from the following Table :

Height in inches :	57	59	61	62	63	64	65	66	67	69
No. of persons :	3	5	7	10	20	22	24	5	2	2

Frequencies given at page 155, in column (1) are grouped by *two*'s in column (2) and (3) and then by *three*'s in columns (4), (5) and (6). The maximum frequency in each column is marked by **Bold Type.** We do not find any fixed point having maximum frequency but changes with the change of grouping. In the following Table, the sizes of maximum frequency in respect of different columns are arranged.

Grouping Table

Height in inches	Frequency					
	(1)	(2)	(3)	(4)	(5)	(6)
57	3			15		
59	5		12		22	
61	7	17				37
62	10		30	52		
63	20	42			66	
64	22		46			51
65	24	29		31		
66	5		7		9	
67		4				
69	2					

Analysis Table

Column	Size of item having maximum frequency				
1				65	
2		63	64		
3			64	65	
4	62	63	64		
5		63	64	65	
6			64	65	66
No. of times	1	3	5	4	1

From the above Table, we find 64 is the size of the item which is most frequented. The mode is, therefore, located at 64.

Note. At a glance from column (1) one might think that 65 is the mode since it contains maximum frequency. This impression is corrected by the process of grouping. So it is not advisable to locate the mode merely by inspection.

(C) *For Continuous Series.*

By inspections or by preparing Grouping Table and Analysis Table, ascertain the modal class. Then to find the exact value of mode, apply the following formula :

$$\text{Mode} = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i,$$

where l = lower class-boundary of modal class,

f_1 = frequency of modal class,

f_0 = frequency of the class preceding the modal class,

f_2 = frequency of the class succeeding the modal class,

i = size of class-interval of modal class.

Example.

Calculate mode from the following data :

Marks	*No. of students*	*Marks*	*No. of students*
above 10	59	above 50	18
,, 20	54	,, 60	8
,, 30	46	,, 70	0
,, 40	34		

We are to convert the cumulative frequency distribution into a simple frequency distribution.

Marks	*No. of students*
10—20	5
20—30	8
30—40	12
40—50	16
50—60	10
60—70	8

The modal class is (40—50), since the max. frequency is 16. Here, $l=40, f_0=12, f_1=16, f_2=10, i=10$

$$\therefore \quad \text{Mode} = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i = 40 + \frac{16-12}{32-12-10} \times 10$$

$$= 40 + \frac{4}{10} \times 10 = 40 + 4 = 44 \text{ marks.}$$

Location of Mode Graphically.

In case of the Frequency Distribution, Mode can be located graphically.

Draw a histogram of the data given. In the inside of the modal class-bar, draw two lines diagonally starting from each upper corner of the bar to upper corner of the adjacent bar (as shown in the next figure). Now draw a perpendicular from the point of intersection of the diagonal lines to the X-axis. The point on the X-axis is read off, which gives the modal value.

Example.

The monthly profits in rupees of 100 shops are distributed as follows :

Profits per shop :

	0—100	100—200	200—300	300—400	400—500	500—600
No. of shops :	12	18	27	20	17	6

Draw the histogram to the data and hence find the modal value. Check this value by direct calculation. [I.C.W.A. Jan. 1964]

Histogram showing the distribution of profits

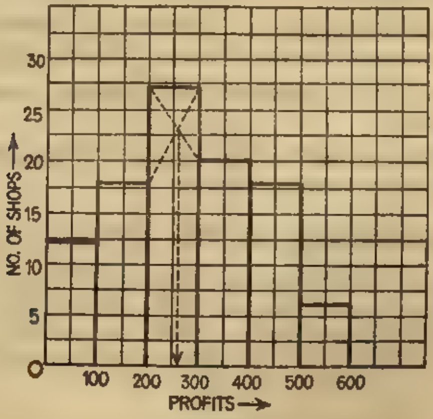


Fig. 34

From the graph, Mode is found to be Rs. 256 (app.).

Now for direct calculation, we find modal class is (200—300), since the class has got the highest frequency.

Again $l=200, f_0=18, f_1=27, f_2=20, i=100$

$$\therefore \quad \text{Mode}=l+\frac{f_1-f_0}{2f_1-f_0-f_2}\times i=200+\frac{27-18}{54-18-20}\times 100$$

$$=200+\frac{9}{16}\times 100=200+56{\cdot}25=\text{Rs. } 256{\cdot}25.$$

Calculation of Mode when class-intervals are unequal : If the class-intervals are unequal, then we are to make them equal, having frequencies adjusted. Then the formula for computing the value of mode is te be applied.

Advantages and Disadvantages of Mode.

Advantages :

(i) It can often be located by inspection.

(ii) It is not effected by extreme values. It is often a really typical value.

(iii) It is simple and precise. It is an actual item of the series except in a continuous series.

(iv) Mode can be determined graphically, unlike Mean.

Disadvantages :

(i) It is unsuitable for algebraic treatment.

(ii) When the number of observations is small, the Mode may not exist, while the Mean and Median can be calculated.

(iii) The value of Mode is not based on each and every item of series.

(iv) It does not lead to the aggregate, if the Mode and the total number of items are given.

Empirical Relationship between Mean, Median and Mode

A distribution in which the values of Mean, Median and Mode coincide, is known symmetrical and if the above values are not equal then the distribution is said asymmetrical or skewed. In a moderately skewed, there is a relation amongst Mean, Median and Mode which is as follows :

$$\textit{Mean}-\textit{Mode}=3\ (\textit{Mean}-\textit{Median}).$$

If any two values are known, we can find the other.

Example.

In a moderately asymmetrical distribution the Mode and Mean are 32·1 and 35·4 respectively. Calculate the Median.

[Delhi, B. Com. 1966]

From the relation, we find 3 Median = 2 Mean + Mode

or, $3 \text{ Median} = 2 \times 35{\cdot}4 + 32{\cdot}1 = 70{\cdot}8 + 32{\cdot}1 = 102{\cdot}9$

$\therefore \quad \text{Median} = 34{\cdot}3.$

Which Average is to apply ?

For all circumstances, no one average can be regarded as best.

A.M. should be avoided in cases of skewed distributions, open-end intervals, for averaging speeds and for extreme items.

G.M. is to applied for construction of index numbers, for computing average rates of increase or decrease.

H.M. is useful for finding rates, time, etc.

Median is the best average in open and grouped distributions, in case of price or income distributions.

Mode is a particularly useful average for discrete series, *i.e.*, number of persons wearing a given size of shoe or number of children per household. For a very large frequency, Mode is suited best.

More Examples.

1. The following data relate to the weights of 90 persons. You are required to form a frequency distribution with class-interval 10 pounds like 100—110, 110—120, etc., and hence compute the Mean, Median, Quartiles and Mode.

134	156	121	159	158	157	112	125	124
136	122	127	167	113	107	152	140	137
103	175	151	139	127	167	198	155	144
136	110	162	119	146	153	142	126	173
145	130	169	117	141	144	165	156	166
177	135	138	168	142	116	195	146	109
176	120	147	133	132	157	143	141	137
154	178	182	135	186	192	170	162	148
140	155	115	147	187	147	129	142	150
146	149	128	160	138	104	181	131	148

Frequency distribution of Weights and Computations of Mean, Median, Quartiles and Mode.

Weight lb	f	Mid-value x	d	$d' = d/10$	fd'	Cum. freq.
(1)	(2)	(3)	(4)	(5)	(6) = (2) × (5)	
100—110	4	105	−40	−4	−16	4
110—120	7	115	−30	−3	−21	11
120—130	10	125	−20	−2	−20	21
130—140	15	135	−10	−1	−15	36
140—150	19	145	0	0	0	55
150—160	13	155	10	1	13	68
160—170	9	165	20	2	18	77
170—180	6	175	30	3	18	83
180—190	4	185	40	4	16	87
190—200	3	195	50	5	15	90 (= N)
Total	90	—	—	—	8	

Let A (assumed Mean) = 145.

$$\text{A.M.} = \text{A} + \frac{\Sigma fd'}{\Sigma f} \times i = 145 + \frac{8}{90} \times 10 = 145 + {}^{\cdot}89 = 145{}^{\cdot}89 \text{ lbs.}$$

Median = size of $\frac{N}{2}$th item = size of $\frac{90}{2}$th item

= size of 45th item. Median lies in the class (140—150).

Here, $l_1 = 140$, $l_2 = 150$, $f_1 = 19$, $m = 45$, $c = 36$.

$$\therefore \quad \text{Median} = l_1 + \frac{l_2 - l_1}{f_1}(m - c) = 140 + \frac{150 - 140}{19}(45 - 36)$$

$$= 140 + \frac{10}{19} \times 9 = 140 + 4{}^{\cdot}74 = 144{}^{\cdot}74 \text{ lbs.}$$

Q_1 (1st quartile) = size of $\frac{N}{4}$th item = size of 22·5th item.

Now, Q_1 lies in the class (130—140)

Here, $l_1 = 130$, $l_2 = 140$, $f_1 = 15$, $q = 22{}^{\cdot}5$, $c = 21$.

$$\therefore \quad Q_1 = l_1 + \frac{l_2 - l_1}{f_1}(q - c) = 130 + \frac{140 - 130}{15}(22{\cdot}5 - 21)$$

$$= 130 + \frac{10}{15} \times 1{\cdot}5 = 130 \text{ lbs.}$$

Q_3 (3rd quartile) = size of $\frac{3N}{4}$th item = size of 67·5th item.

Now Q_3 lies in the class (150 − 160).

Here $l_1 = 150$, $l_2 = 160$, $f_1 = 13$, $q = 67{\cdot}5$, $c = 55$

$$\therefore \quad Q_3 = l_1 + \frac{l_2 - l_1}{f_1}(q - c) = 150 + \frac{160 - 150}{13}(67{\cdot}5 - 55)$$

$$= 150 + \frac{10}{13} \times 12{\cdot}5 = 150 + 9{\cdot}62 = 159{\cdot}62 \text{ lbs.}$$

The modal class is (140—150).

Here, $l = 140$, $f_0 = 15$, $f_1 = 19$, $f_2 = 13$, $i = 10$

$$\therefore \quad \text{Mode} = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i \quad = 140 + \frac{19 - 15}{38 - 15 - 13} \times 10$$

$$= 140 + \frac{4}{10} \times 10 = 140 + 4 = 144 \text{ lbs.}$$

2. An incomplete distribution is given below :

Variable	*Frequency*
10—20	12
20—30	30
30—40	?
40—50	65
50—60	?
60—70	25
70—80	18
Total	229

You are given that the median value is 46.

(a) Using the median formula fill up the missing frequencies,

(b) calculate the Arithmetic mean of the completed Table.

[C. A. May 1968]

Let the frequency of the class (30—40) $=f_1$
and " " " " " (50—60) $=f_2$.
Now $12+30+f_1+65+f_2+25+18=229$
or, $f_1+f_2=229-150=79$.

Median $=$ size of $\frac{N}{2}$th item $=$ size of $\frac{229}{2}$th item

$=$ size of 114·5th item.

Now median lies in the class (40—50), since median $=46$.
Here, $l_1=40$, $l_2=50$, $f=65$, $m=114{\cdot}5$, $c=42+f_1$.

From, median $=l_1+\frac{l_2-l_1}{f}(m-c)$, we get,

$$46=40+\frac{50-40}{65}\{114{\cdot}5-(12+30+f_1)\}$$

$$\text{or,}\quad 46=40+\frac{10}{65}(72{\cdot}5-f_1)$$

$$\text{or,}\quad 46-40=\frac{10}{65}(72{\cdot}5-f_1)$$

$$\text{or,}\quad 6=\frac{10(72{\cdot}5-f_1)}{65}$$

$$\text{or,}\quad f_1=33{\cdot}5=34 \text{ (app.)}$$

$$\therefore\quad f_2=79-f_1=79-34=45$$

$$\therefore\quad f_1=34 \quad\text{and}\quad f_2=45.$$

For Computation of Arithmetic Mean

Variable	*Mid-value* x	f	d	$d'=d/10$	fd'
10—20	15	12	-30	-3	-36
20—30	25	30	-20	-2	-60
30—40	35	34	-10	-1	-34
40—50	45	65	0	0	0
50—60	55	45	10	1	45
60—70	65	25	20	2	50
70—80	75	18	30	3	54
Total	—	229	—	—	19

Let A = 45

$$\therefore \quad \text{A.M.} = \text{A} + \frac{\Sigma fd'}{\Sigma f} \times i = 45 + \frac{19}{229} \times 10 = 45 + {}^{\cdot}83 = 45{}^{\cdot}83 \text{ (app.)}.$$

3. The Table below gives the diastolic blood pressure of 250 men. The readings were made to the nearest millimetre and the central value of each group is given :

Blood pressure (*mm*) :	60	65	70	75	80	85	90	95
Number of men :	4	5	31	39	114	30	25	2

Calculate from the data the mean and median.

[I.C.W.A. July 1970]

For calculation of median we are to form the class-boundaries from the mid-values given. The common difference between the mid-values, indicates the class-boundaries will be of equal width.

Calculation of Mean and Median

Mid-values	*Class-boundaries*	f	d	$d' = \frac{d}{5}$	fd'	*Cum. frequency*
60	57·5—62·5	4	−20	−4	−16	4
65	62·5—67·5	5	−15	−3	−15	9
70	67·5—72·5	31	−10	−2	−62	40
75	72·5—77·5	39	−5	−1	−39	79
80	77·5—82·5	114	0	0	0	193
85	82·5—87·5	30	5	1	30	223
90	87·5—92·5	25	10	2	50	248
95	92·5—97·5	2	15	3	6	250 (= N)
Total	—	250	—	—	−46	

Let A = 80, $\text{A.M.} = \text{A} + \dfrac{\Sigma fd'}{\Sigma f} \times i$

$$= 80 + \frac{(-46)}{250} \times 5 = 80 - {}^{\cdot}92 = 79{}^{\cdot}08 \; mm.$$

Median = size of $\frac{\text{N}}{2}$th item = size of $\frac{250}{2}$(= 125)th item.

Median lies in (77·5—82·5). $l_1 = 77·5$, $l_2 = 82·5$, $f_1 = 144$

$$\therefore \quad \text{Median} = l_1 + \frac{l_2 - l_1}{f_1}(m - c)$$

$$= 77·5 + \frac{5}{114}(125 - 79)$$

$$= 77·5 + \frac{5}{114} \times 46 = 77·5 + 2·02 = 79·52 \ mm.$$

4. Given below is the distribution of 140 candidates obtaining marks X or higher in a certain examination (all marks are given in whole numbers) :

X :	10	20	30	40	50	60	70	80	90	100
c.f. :	140	133	118	100	75	45	25	9	2	0

Calculate the mean marks obtained by the candidates.

[I.C.W.A. Inter. June 1975]

For calculating mean at first we are to transfer the cumulative frequency distribution (given in greater than type) in the form of a frequency distribution as follows and hence to apply the usual formula.

Frequency Distribution and Calculation of Mean

Class-intervals	*Frequency* (f)	*Mid.-pt.* (x)	d	$d' = \frac{d}{10}$	fd'
10— 20	7	15	−40	−4	−28
20— 30	15	25	−30	−3	−45
30— 40	18	35	−20	−2	−36
40— 50	25	45	−10	−1	−25
50— 60	30	55	0	0	0
60— 70	20	65	10	1	20
70— 80	16	75	20	2	32
80— 90	7	85	30	3	21
90—100	2	95	40	4	8
Total	140	—	—	—	−53

Let A = 55

$$\text{A.M.} = \text{A} + \frac{\Sigma fd'}{\Sigma f} \times i = 55 + \frac{(-53)}{140} \times 10 = 55 - 3·78 = 51·22 \text{ marks.}$$

5. The numbers 3·2, 5·8, 7·9 and 4·5 have frequencies x, $(x+2)$, $(x-3)$ and $(x+6)$ respectively. If the arithmetic mean is 4·876, find the value of x. [C.U. M. Com. 1973]

Calculation of the value of x

Numbers x	*Frequency* f	fx
3·2	x	$3·2x$
5·8	$x+2$	$5·8x+11·6$
7·9	$x-3$	$7·9x-23·7$
4·5	$x+6$	$4·5x+27·0$
Total	$4x+5$	$21·4x+14·9$

Now A.M. $=\frac{\Sigma fx}{\Sigma f}$

or, $4·876=\frac{21·4x+14·9}{4x+5}$

or, $4·876(4x+5)=21·4x+14·9$

or, $19·504x+24·380=21·4x+14·9$

or, $1·896x=9·480$

or, $x=\frac{9·480}{1·896}=5.$

6. Put the following information into a frequency distribution and obtain the arithmetic mean (assuming the range of salary is Rs. 0—500) :—

For a group of wage-earners, 20%, 40%, 70% and 80% of the wage-earners receive less than Rs. 50, 120, 300 and 350 respectively ; and 5% are receiving Rs. 400 and over.

[C.U. B.A. (Econ.) 1965]

From the question it is clear that balance 15% of the wage-earners will lie in the group of Rs. 350 and less than Rs. 400.

From the cumulative percentage distribution as shown in the first table (below), we form the grouped frequency distribution as shown in the next Table :

Cum. Percentage Distribution

Salary (in Rs.)	*Cum. percentage*
0	0
50	20
120	40
300	70
350	80
400	95
500	100

Frequency Distribution of Wages

Salary (in Rs.)	*Percentage of wage-earners*
0— 50	20
50—120	20
120—300	30
300—350	10
350—400	15
400—500	5
Total	100

Taking percentage of wage-earners as weights, calculation of arithmetic mean is shown in the Table below :—

Salary (in Rs.)	*Mid-value* x	f	d	$d'=d/5$	fd'
0— 50	25	20	−300	−60	−1200
50—120	85	20	−240	−48	−960
120—300	210	30	−115	−23	−690
300—350	325	10	0	0	0
350—400	375	15	50	10	150
400—500	450	5	125	25	125
Total	—	100	—	—	−2575

Let = 325

$$\therefore \quad \text{A.M.} = \text{A} + \frac{\Sigma fd'}{\Sigma f} \times i = 325 + \frac{(-2575)}{100} \times 5$$

$$= 325 - 128{\cdot}75 = \text{Rs. } 196{\cdot}25.$$

7. (a) The following frequency distribution is with regard to weight in gm. of mangoes of a given variety. If mangoes of weight less than 443 gms. be considered unsuitable for foreign market what is

the percentage of the yield suitable for it ? Assume the given frequency distribution to be typical of the variety.

Weight in gms. :	410—419	420—429	430—439	440—449	450—459	460—469	470—479	Total
Frequen. :	14	20	42	54	45	18	7	200

(b) Draw an ogive of 'more than' type on the data of the above question. Deduce from it the median of the distribution.

[I.C.W.A. July 1968]

(a)

Weight (gm.)	*Frequency*	*Cumulative frequency*
410—419	14	14
420—429	20	34
430—439	42	42
440—449	54	130
450—459	45	175
460—469	18	193
470—479	7	200
Total	200	

Let x = total number of mangoes of weight less than 443 gms. which lies in (440—449).

Now, applying the interpolation formula, we find (taking lower and upper class-boundaries).

$$443 = 439{\cdot}5 + \frac{449{\cdot}5 - 439{\cdot}5}{54}(x - 76)$$

$$\text{or,} \quad 443 - 439{\cdot}5 = \frac{10}{54}(x - 76)$$

$$\text{or,} \quad 3{\cdot}5 = \frac{10x - 760}{54}$$

$$\text{or,} \quad 10x - 760 = 3{\cdot}5 \times 54 = 189$$

$$\therefore \quad x = 94{\cdot}9 = 95 \text{ (app.)}$$

Number of suitable mangoes = 200 − 95 = 105

∴ Percentage of total yield (suitable) $= \frac{105}{200} \times 100 = 52{\cdot}5\%$.

b)

Class-boundaries	*Frequency*	*Cum. freq.* 'more than'
409·5—419·5	14	200
419·5—429·5	20	186
429·5—439·5	42	166
439·5—449·5	54	124
449·5—459·5	45	70
459·5—469·5	18	25
469·5—479·5	7	7

Ogive 'more than' type is to be drawn, from the above Table (drawing is left to the students), and hence estimate median by usual process.

[For check, median $= 439{\cdot}5 + \frac{10}{54}(100 - 70) = 445{\cdot}06$

$= 445$ (app.) gms.]

8. Three groups of observations contain 8, 7 and 5 observations. Their geometric means are 8·52, 10·12 and 7·75 respectively. Find the geometric mean of the 20 observations in the single group formed by pooling the three groups.

[I.C.W.A. Dec. '76]

Here, $n_1 = 8$, $n_2 = 7$, $n_3 = 5$, $G_1 = 8{\cdot}52$, $G_2 = 10{\cdot}12$, $G_3 = 7{\cdot}75$.

G. M. (G) of the composite group is

$G = \sqrt[N]{G_1^{n_1}.G_2^{n_2}.G_3^{n_3}}$, where G_1, G_2, G_3 are G.M. of groups having n_1, n_2, n_3 observations respectively.

or, $\log G = \frac{1}{N}\left(n_1 \log G_1 + n_2 \log G_2 + n_3 \log G_3\right)$

$$\text{or,}\quad \log G = \frac{1}{n_1+n_2+n_3}\left(n_1 \log G_1 + n_2 \log G_2 + n_3 \log G_3\right)$$

$$= \frac{1}{8+7+5}\left(8 \log 8{\cdot}52 + 7 \log 10{\cdot}12 + 5 \log 7\;75\right)$$

$$= \frac{1}{20}\left(8 \times {\cdot}9304 + 7 \times 1{\cdot}0052 + 5 \times {\cdot}8893\right)$$

$$= \frac{1}{20}\left(7{\cdot}4432 + 7{\cdot}0364 + 4{\cdot}4465\right)$$

$$= \frac{1}{20} \times 18{\cdot}9261$$

$$= 0{\cdot}9463 = \log 8{\cdot}837$$

$$\therefore \quad \text{G.M.} = 8{\cdot}837 = 8{\cdot}84.$$

EXERCISE 6

1. Point out the advantages and disadvantages of the chief kinds of averages used in statistics.

2. What is the difference between simple and weighted average ? Explain the circumstances under which the latter should be used in preference to the former.

3. Define the different measures of central tendency explaining how each of them can be computed for a given frequency distribution.

4. In each of the following cases, explain whether the description applies to the mean, median or both :

(i) Can be calculated from a frequency distribution with open-end classes.

(ii) The values of all items are taken into consideraton in the calculation.

(iii) The values of extreme items do not influence the average.

(iv) In a distribution with a single peak and moderate skewness to the right, it is closer to the concentration of the distribution.

[C. A. Nov. 1965] (*Ans.* Median ; both ; Median ; Median)

5. State and prove the properties of Geometric Mean.

6. The following are the monthly salaries in rupees of 20 employees of a firm ;

130	62	145	118	125	76	151	142	110	98
65	116	100	103	71	85	80	122	132	95

The firm gives bonuses of Rs. 10, 15, 20, 25 and 30 for individuals in the respective salary groups exceeding Rs. 60 but not exceeding Rs. 80 ; exceeding Rs. 80 but not exceeding Rs. 100 and so on upto exceeding Rs. 140 but not exceeding Rs. 160. Find the average bonus paid per employee. [C. A. Nov. 1964] (*Ans.* Rs. 19)

7. Calculate the value of median, mode and two quartiles from the following data :

Age	*No. of persons*	*Age*	*No. of persons*
20—25	50	40—45	100
25—30	70	45—50	120
30—35	100	50—55	70
35—40	180	55—60	60

[I.C.W.A. 1966] (*Ans.* Med = 40, Mode = 38·64, $Q_1 = 34$, $Q_3 = 47·08$)

8. The frequency distribution below gives the cost of production of sugarcane in different holdings. Obtain the Arithmetic Mean.

Cost	*Frequency*	*Cost*	*Frequency*
2—6	1	18—	52
6—	9	22—	36
10—	21	26—	19
14—	47	30—34	3

(*Ans.* 19·21)

9. Find the median height of Indian adult males from the following frequency distribution :

Height (cm.)	*Frequency*
144·55—149·55	1
149·55—154·55	3
154·55—159·55	24
159·55—164·55	58
164·55—169·55	60
169·55—174·55	27
174·55—179·55	2
179·55—184·55	2

(*Ans.* 164·76 *cm.*)

10. An incomplete frequency distribution is given below :

Height (inches) :	5·1—6·0	6·1—7·0	7·1—8·0	8·1—9·0	9·1—10·0	10·1—11·0	11·1—12·0
No. of plants :	3	8	27	?	17	11	9

It is known that the median height of a plant is 8·53 inches. Calculate the missing frequency. [I.C.W.A. Jan. 1972] (*Ans.* 25)

11. The Arithmetic mean calculated from the following frequency distribution is known to be 67·45 inches. Find the value of f_3.

Height (inches) :	60—62	63—65	66—68	69—71	72—74
Frequency :	15	54	f_3	81	24

[I.C.W.A. July 1971] (*Ans.* 126)

12. Find the Arithmetic mean and the years in which the modal point and the median fall from the following data :

The No. of persons killed in accident in the coal mines in India

Year :	1951	'52	'53	'54	'55	'56	'57	'58	'59
No. :	319	353	330	429	309	259	182	420	212

Find also where the qaurtiles will lie. [C.U.B. Com. (II) 1966]
(*Ans.* : Mean = 312·56 ; 1954 ; Q_1 in 1954 ; 1951 ; Q_3 in 1952)

13. Comment on the performance of the students of the three universities given below using simple and weighted averages :

University :	*Bombay*		*Calcutta*		*Madras*	
Course of Study	*% of pass*	*No of students* (in hundreds)	*% of pass*	*No of students* (in hundreds)	*% of pass*	*No of students* (in hundreds)
M. A.	71	3	82	2	81	2
M. Com.	83	4	76	3	76	3·5
B. A.	73	5	73	6	74	4·5
B. Com.	74	2	76	7	58	2
B. Sc.	65	3	65	3	70	7
M. Sc.	66	3	60	7	73	2

[C.A. Nov. 1970]

(*Ans.* : Simple average : 72 ; 72 ; 72
Wt. average : 72·55 ; 70·61 ; 72·55)

14. The table given below has been constructed from data obtained from a factory showing the distribution of the number of processed articles per day per person and the rate of payment :

Daily no. of articles processed per person	*No. of persons processing*	*Rate of payment per article processed* (P)
80—99	12	3·1
100—119	63	3·2
120—139	87	3·3
140—159	56	3·4
160—169	3	3·5

Calculate the rate of payment per person per article processed.

[I.C.W.A. July 1964] (*Ans.* 3·301 P)

15. There are two branches of an establishment employing 100 and 80 persons respectively. If the arithmetic means of the monthly salaries paid by two branches are Rs. 275 and Rs. 225 respectively, find the arithmetic mean of the salaries of the employees of the establishment as a whole. [C.A. Nov. 1963] (*Ans.* Rs. 252·78)

16. The mean age of a group of 100 children was 9·35 years. The mean age of 25 of them was 8·75 years and that of another 65 was 10·51 years. What was the mean age of the remainder ?

[C.U. M.Com. 1965] (*Ans.* 3·31 yrs.)

17. From an income distribution of a group of mean 20% of men have income below Rs. 30, 35% below Rs. 70, 60% below Rs. 150 and 80% below Rs. 250. The first and third quartiles are Rs. 50 and Rs. 170.

Put the above information in a cumulative frequency distribution and find the median. [C.U. M.Com. 1966] (*Ans.* Rs. 118)

18. For a certain group of 'saree' weavers of Varanasi, the median and quartiles of earnings per week are Rs. 44·30, Rs. 43·00 and Rs. 45·90 respectively. 10% of the group earn under Rs. 42 per week and 13% earn Rs. 47 and over and 6% Rs. 48 and over. The range of earnings per week is Rs. 40—Rs. 50. Put the data into a frequency distribution. [C.U. B.A. (Econ.) 1970]

19. Find graphically the median value from the following data on yield of grain in pounds per 1/500 acre :

Yield	*No. of plots*	*yield*	*No. of plots*
2·7—2·9	4	4·1—4·3	69
2·9—3·1	15	4·3—4·5	59
3·1—3·3	20	4·5—4·7	35
3·3—3·5	47	4·7—4·9	10
3·5—3·7	63	4·9—5·1	8
3·7—3·9	78	5·1—5·3	4
3·9—4·1	88		

Determine the modal value from its approximate relationship with mean and median. [I.A.S. 1962] (*Ans.* 3·95)

20. In a sample survey of 60 workers' families living in a factory area, the following data were obtained, as regards the number of members in the families. Form a frequency distribution and find the mean and median family size.

5	11	4	6	3	10	5	7	9	6
6	2	3	7	8	6	4	3	6	5
9	5	6	4	7	1	5	8	6	2
6	4	8	7	5	12	4	7	10	6
7	8	3	6	7	5	5	8	6	4
6	11	5	2	6	9	7	3	7	5

[I.C.W.A. July 1969]

(*Ans.* Freq. : 1, 3, 5, 6, 10, 13, 9, 5, 3, 2, 2, 1 ; Mean = 5·97 ; Median = 6)

21. The weekly wages earned by the hundred workers of a factory are set out in the following table :

Weekly wages (Rs.) :	12·5—17·5	17·5—22·5	22·5—27·5	27·5—32·5	32·5—37·5	37·5—42·5	42·5—47·5	47·5—52·5	52·5—57·5
No. of workers :	12	16	25	14	13	10	6	3	1

Calculate the three quartiles of the above distribution, taking $\frac{n}{4}$, $\frac{2n}{4}$ and $\frac{3n}{4}$ as their ranks. [C.A. Nov. 1963]

(*Ans.* $Q_1 = 21·56$; $Q_2 = 26·9$; $Q_3 = 35·58$)

22. The following are the marks obtained by a batch of 20 students in a certain class-test in English and Mathematics :

Roll No.	1	2	3	4	5	6	7	8	9	10
Marks in English	53	54	52	32	30	60	47	46	35	28
Marks in Mathematics	58	55	25	32	26	85	44	80	33	72
Roll No.	11	12	13	14	15	16	17	18	19	20
Marks in English	25	42	33	48	72	51	45	33	65	29
Marks in Mathematics	10	42	15	46	50	64	39	38	30	36

In which subject is the level of knowledge of the students higher ? [Gorakhpur, B. Com. 1966] (*Ans.* Median (Eng.) = 45·5, Median (Math.) = 41·5 knowledge of English is higher).

23. Find the median and mode from the following table :

No. of days absent	*No. of students*	*No. of days absent*	*No. of students*
less than 5	29	less than 30	644
,, ,, 10	224	,, ,, 35	650
,, ,, 15	465	,, ,, 40	653
,, ,, 20	582	,, ,, 45	55
,, ,, 25	634		

[C.A. May 1965] (*Ans.* Median = 12·75, Mode = 11·35)

24. From the results of the two colleges A and B, given below, state which of them is better and why.

Class	*A-College*		*B-College*	
	Appeared	Passed	Appeared	Passed
M. A.	30	25	100	80
M. Com.	50	45	120	95
B. A.	200	150	155	70
B. Com.	120	75	85	50
Total	400	295	455	295

[Lucknow, B. Com. 1949] (*Ans.* A-College)

25. Find the mean, median and modal ages of married women at first child-births.

Age at the birth of 1st child :	13	14	15	16	17	18	19	20	21	22	23	24	25
No. of married women :	37	162	343	390	256	433	161	355	65	85	49	46	40

[I.C.W.A. Jan.] (*Ans.* 17·72 ; 18 ; 18)

26. Frequency distribution of weekly wages of 500 workmen :

Weekly wages (Rs.) :	11—15	16—20	21—25	26—35	36—45	46—60	61—75	76—100	Total
Freq. :	2	23	86	154	120	75	33	7	500

Draw an ogive of this distribution and use it to find (a) the median wage, (b) the wage limits of the central 50% of the wage-earners and (c) the percentage of workmen earning more than Rs. 32·50 per week. [I.C.W.A. Jan. 1967]

[*Ans.* (a) Rs. 34·53, (b) from Q_1 = Rs. 26·41 to Q_3 = Rs. 44·67, (c) 56%]

27. The following table shows the age-distribution of heads of families in a certain country during the year 1957. Find the median, the third quartile, and the second decile of the distribution. Check the results by the graphical method.

Age of head of family (years)	*Number* (million)
under 25	2·3
25—29	4·1
30—34	5·3
35—44	10·6
45—54	9·7
55—64	6·8
65—74	4·4
above 74	1·8
Total	45·0

[I.C.W.A. Jan. 1973] (*Ans.* 44·71 yrs. ; 57·1 yrs. ; 31·95 yrs.)

28. Below is given the frequency distribution of weights of a group of 60 students in a class in a school :

Wt. (Kg.) :	30—34	35—39	40—44	45—49	50—54	55—59	60—64
No. of Students :	3	5	12	18	14	6	2

(*a*) Draw histogram for this distribution and find the modal value.

(*b*) Prepare the (1) cumulative frequency (both less than and more than types) distribution and (2) represent them graphically on the same graph paper. Hence find the (3) median, (4) quartile deviation.

(*c*) With the modal and the median values as obtained in (*a*) and (*b*), use an appropriate empirical formula to find the arithmetic mean of this distribution.

(*d*) If students obtaining marks below 40 are eliminated from the frequency distribution, what will be the revised mean ? Calculate the mean of the two rejected classes only and use the result obtained in (*c*). [I.C.W.A. June 1976] (*Ans.* 47·5 Kg. ; 47·3 Kg. ; 4·8 Kg. ; 47·2 Kg. ; 49·1 Kg.

29. Explain what is meant by central tendency of data. What are the common measures of central tendency ? [I.C.W.A. June 76]

30. Point out the merits and demerits of the mean, the median and the mode as measures of central tendency of numerical data. [I.C.W.A. Dec. '76]

31. Form an ordinary frequency table from the following cumulative distribution of marks obtained by 22 students and calculate (i) A.M., (ii) Median and (iii) Mode.

Marks		*No. of students*
Below	10	3
,,	20	8
,,	30	17
,,	40	20
,,	50	22

[I.C.W.A. June 77] (*Ans.* (i) 23·18 marks, (ii) 23·33 marks, (iii) 24 marks)

32. (*a*) Given the following frequency distribution, calculate the mean :

Monthly wages (in Rs.)	*No. of workers*
12·5—17·5	2
17·5—22·5	22
22·5—27·5	10
27·5—32·5	14
32·5—37·5	3
37·5—42·5	4
42·5—47·5	6
47·5—52·5	1
52·5—57·5	1
Total	63

(*b*) Draw cumulative frequency diagram (less than type) of the above frequency distribution and hence determine the median wages. [I.C.W.A. Dec. '77] (*Ans.* Rs. 28·25 ; Rs. 25·75)

33. The following table gives the Vickers Hardness numbers of 20 shell cases :

66·3	61·3	62·7	60·4	60·2
64·5	66·5	62·9	61·5	67·8
65·0	62·7	62·2	64·8	65·8
62·2	67·5	67·5	60·9	63·8

Draw the cumulative frequency diagram of these numbers (Either less than type or more than type need be drawn).

Determine the range, upper and lower quartiles, inter-quartile range and median. Indicate the quartiles and the median on the cumulative frequency diagram.

[I.C.W.A. June '78] (*Ans.* 7·6 ; 65·8 ; 61·5 ; 4·3 ; 63·35)

34. The expenditure of 1000 families is given as under :

Expenditure : (in Rs.)	40—59	60—79	80—99	100—119	120—139
No. of families :	50	?	500	?	50

The median and mean for the distribution are both Rs. 87·50 paise respectively. Calculate the missing frequencies.

[I.C.W.A. June 78] (*Ans.* 250 ; 150)

35. An aeroplane flies around a square the sides of which measure 100 Kms. each. The aeroplane covers at a speed of 100 Kms. per hour the first side, at 200 Kms. per hour the second side, at 300 Kms. per hour the third side and at 400 Kms. per hour the fourth side. Use the correct mean to find the average speed round the square.

[I.C.W.A. June 78] (*Ans.* 192 Kms. per hour)

36. (*a*) Explain what is mean by central tendency of data. What are the common measures of central tendency ?

(*b*) Given below the frequency distribution of carbon content (present) in 150 determinations on a certain mixed powder.

Present Carbon	*Frequency*
4·0—4·1	1
4·2—4·3	2
4·4—4·5	7
4·6—4·7	20
4·8—4·9	25
5·0—5·1	30
5·2—5·3	10
5·4—5·5	25
5·6—5·7	30

—Compute the arithmetic mean, median.

[I.C.W.A. Dec. '78] (*Ans.* 5·118 ; 5·083)

37. Below is given the frequency distribution of marks in Mathematics obtained by 100 students in a class :

Marks	*No. of students*
20—29	8
30—39	10
40—49	25
50—59	31
60—69	11
70—79	12
80—89	2
90—99	1
Total	100

Draw the ogive for this distribution and use it to determine the median. [I.C.W.A. June 79] (*Ans.* 51·8 marks)

38. (*a*) In a certain country, the age-distribution of women in 1947 is as follows :

Years of age	*Millions*
under 10	3·75
10 and under 20	3·30
20 „ „ 30	3·65
30 „ „ 40	3·95
40 „ „ 50	3·65
50 „ „ 60	3·15
60 „ „ 70	2·45
70 „ „ 80	2·10

Calculate the (i) A.M., (ii) Mode, (iii) Median, and (iv) First Quartile.

(*b*) Prove that for any two real quantities A.M. > G.M.>H.M.

[I.C.W.A. June 79] (*Ans.* (i) 36·61 yrs., (ii) 35 yrs., (iii) 35·82 yrs., (iv) 18·33 yrs.)

39. Calculate the mean and the median from the following data :

Weekly wages (Rs.)	*Number of workers*
Below 10	8
" 20	18
" 30	45
" 40	90
" 50	113
" 60	120

[C. U. B. Com. (Hons.) 1980] (*Ans.* Rs. 32·17 ; Rs. 33·33)

40. The frequency distribution of weekly wages in a certain factory is as follows :

Weekly wages (Rs.)	*Number of workers*
23—27	2
28—32	6
33—37	9
38—42	14
43—47	32
48—52	16
53—57	12
58—62	6
63—67	2
68—72	1
Total	100

Draw the ogive (less than or more than type) of this distribution and find from the ogive (i) the first quartile, (ii) the median, and (iii) the third quartile.

[I.C.W.A. Dec. 1979] (*Ans.* Rs. 40 ; Rs. 42 ; Rs. 51)

41. Find the median and mode for the following distribution :

No. of days absent	*No. of students*
Less than 5	30
,, ,, 10	225
,, ,, 15	465
,, ,, 20	580
,, ,, 25	634
,, ,, 30	644
,, ,, 35	650
,, ,, 40	653
,, ,, 45	655

[I.C.W.A. Dec. 1979] (*Ans.* 12·14 days ; 11·32 days)

8

DISPERSION

Introduction.

The various measures of central tendency give us one single figure to represent the entire data. But the average, as we have seen, has its own limitations. There are number of series whose averages may be identical, but differ from each other in many ways. In such cases further statistical analysis of the data is necessary to study these differences. Measures of dispersion help us to study the characteristics, *i.e.*, the extent to which the items (or observations) differ from one another and from central value.

Suppose there are three series of 5 items, each as follows :

A :	50	50	50	50	50,	total = 250,	mean = 50
B :	48	45	52	50	55,	total = 250,	mean = 50
C :	2	110	40	30	68,	total = 250,	mean = 50

In A, the values of all the items are same and not deviated (or scattered) from the mean. There is no dispersion.

In B, only one item is perfectly represented by mean, the other, items are not very much scattered as the minimum value is 48 and the maximum is 55.

In C, not a single figure is represented by mean and the items vary widely. The dispersion is very much, in comparison with B. Obviously, the average does not satisfactorily represent the items.

For correct analysis of these series, we are to study something more than their averages. From above, it is clear that a study regarding the deviations about an average should be accounted for. This kind of deviation is known as dispersion.

A *measure of dispersion* is designed to state the extent to which individual observations (or items) vary from their average. Here we shall account only the amount of variation (or its degree) and not the direction (which will be discussed later on in connection with skewness).

Usually, the deviations of the observations from their average (mean, median or mode) are found out, then the average of these deviations is taken to represent the dispersion of a series. This is why measures of dispersion are known as *Averages of the second order.* We have seen earlier, mean, median and mode, etc. are all *Averages of the first order.*

Types.

Measures of dispersion are mainly of two types :

(A) *Absolute measures,* (B) *Relative measures.*

(A) Absolute measures are of four types :

(i) Range

(ii) Quartile deviation (or Semi-interquartile range)

(iii) Mean deviation (or Average deviation)

(iv) Standard deviation

(B) Among the Relative measures we find the following types :

(i) Coefficient of quartile deviation

(ii) Coefficient of dispersion

(iii) Coefficient of variation

Absolute and Relative measures : If we calculate dispersion of a series, say, marks obtained by students in absolute figures, then dispersion will be also in the same unit (*i.e.*, marks). This is Absolute dispersion. If again, dispersion is calculated as a ratio (or percentage) of the average, then it is Relative dispersion.

Range.

For a set of observations, range is the difference between the extremes, *i.e.*,

$$Range = Maximum\ Value - Minimum\ Value.$$

Example.

The marks obtained by 6 students were 24, 12, 16, 11, 40, 42. Find the Range. If the highest mark is omitted, find the percentage change in range.

Here maximum mark = 42, minimum mark = 11.

$\therefore$ Range = 42 − 11 = 31 marks.

If again, the highest mark 42 is omitted, then amongst the remainings, maximum mark is 40.

So range (revised) = 40 − 11 = 29 marks.

Change in range = 31 − 29 = 2 marks.

$\therefore$ reqd. percentage change $= \frac{2}{6} \times 100 = 33{\cdot}33\%$.

Note. Range and other absolute measures of dispersion are to be expressed in the same unit in which observations are expressed.

Advantages and Disadvantages of Range.

Advantages : Range is easy to understand and is simple to compute.

Disadvantages : It is very much effected by the extreme values. It does not depend on all the observations, but only on the extreme values. Range cannot be computed in case of open-end distribution.

Uses of Range.

It is popularly used in the field of quality control. In stock-market fluctuations, range is used.

Quartile Deviation (Q.D.).

The Quartile Deviation is half of the difference between the upper and lower quartiles.

$$\therefore \quad \textit{Quartile Deviation} = \tfrac{1}{2}(Q_3 - Q_1).$$

By *Inter-quartile range*, we understand the difference between two quartiles (*i.e.*, $Q_3 - Q_1$), and half of this means Semi-interquartile range (*semi* stands for *half*).

Since 50% of the observations lie between two quartiles, as such Inter-quartile range gives a fair measure of variability. Interquartile range also does not depend on all observations, and it is effected by fluctuations.

Quartile Deviations (Q.D.) is an absolute measure of dispersion. If it is divided by average value of two quartiles, we will find *Coefficient of Quartile Deviation* (a relative measure of dispersion).

Symbolically, coefficient of quartile deviations $= \frac{\frac{1}{2}(Q_3 - Q_1)}{\frac{1}{2}(Q_3 + Q_1)} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$.

FOR INDIVIDUAL OBSERVATIONS :

Example. Find the quartile deviation and coefficient of quartile deviation of the following observations :

(Marks) 11, 12, 14, 17, 19, 21, 27, 28, 30, 32, 33.

Here, $n = 11$, and observations are arranged in order.

Q_1 = size of $\frac{n+1}{4}$th item = size of 3rd item = 14 marks.

Q_3 = size of $\frac{3(n+1)}{4}$th item = size of 9th item = 30 marks.

$\therefore$ Quartile Deviation (Q. D.) $= \frac{30-14}{2} = \frac{16}{2} = 8$ marks.

Again, Coefficient of quartile deviation $= \frac{30-14}{30+14} = \frac{16}{44} = \cdot 363$.

FOR DISCRETE SERIES :

Example. Compute coefficient of quartile deviation from the following data :

Wages (Rs.) :	12	14	17	21	27	30	36	Total
No. of workers :	4	6	8	7	12	10	4	51

Cumulative Frequency Distribution

Wages (Rs.)	*f*	*Cum. freq.*
12	4	4
14	6	10
17	8	18
21	7	25
27	12	37
30	10	47
36	4	51 (= N)

Q_1 = size of $\frac{N+1}{4}$th item = size of 13th item = Rs. 17.

Q_3 = size of $\frac{3(N+1)}{4}$th item = size of $\frac{3 \times 52}{4}$th item

= size of 39th item = Rs. 30.

Here Q. D. $= \frac{30-17}{2} = \frac{13}{2}$ = Rs. 6·5.

$$\therefore \quad \text{Coefficient of Q. D.} = \frac{30-17}{30+17} = \frac{13}{47} = \cdot 277.$$

For Continuous Series :

Example. Calculate appropriate measure of dispersion from the following data :

Wages in rupees per week	*No. of wage-earners*
less than 35	14
35—37	62
38—40	99
41—43	18
over 43	7

[I.C.W.A. Jan. 1964]

In the frequency distribution, there are open-end classes, so Q. D. would be the appropriate measure of dispersion.

Cumulative Frequency Distribution

Wages (Rs.)	*f*	Cum. freq.
less than 35	14	14
35—37	62	76
38—40	99	175
41—43	18	193
over 43	7	200(=N)

$$Q_1 = \text{size of } \frac{N}{4}\text{th item} = \text{size of 50th item.}$$

Q_1 lies in the class (34·5—37·5).

$$\therefore \quad Q_1 = 34{\cdot}5 + \frac{(37{\cdot}5 - 34{\cdot}5)}{62}(50-14) = 34{\cdot}5 + \frac{3}{62} \times 36$$

$$= 34{\cdot}5 + 1{\cdot}74 = \text{Rs. } 36{\cdot}24.$$

Q_3 = size of 150th item, Q_3 lies in the class (37·5—40·5)

$$\therefore \quad Q_3 = 37{\cdot}5 + \frac{(40{\cdot}5 - 37{\cdot}5)}{99}(150-76) = 37{\cdot}5 + \frac{3}{99} \times 74$$

$$= 37{\cdot}5 + 2{\cdot}24 = \text{Rs. } 39{\cdot}74.$$

$$\therefore \text{ Quartile Deviation} = \frac{39{\cdot}74 - 36{\cdot}24}{2} = \frac{3{\cdot}50}{2} = \text{Rs. } 1{\cdot}75.$$

Note. Since mid-values of open-end classes cannot be determined, mean deviation and standard deviation cannot be calculated.

Advantages and Disadvantages of Quartile Deviation.

Advantages : It is superior to range as measures of dispersion. In case of open-end distributions, it can be computed. It is not effected by the presence of extreme values.

Disadvantages : Quartile deviation is neither based on all the observations nor is it capable of further algebraic treatment. Its value is much effected by sampling fluctuations. It is not a measure of dispersion, particularly for series in which variation is considerable.

Mean Deviation (or Average Deviation).

The two methods—Range and Quartile Deviation are calculated, based on only two points of a series—extreme values in case of range and quartiles for quartile deviation. They are not based on all the observations. Mean deviation and standard deviation, however, are computed by taking into account all the observations of the series.

Definition.

Mean deviation of a series is the arithmetic average of the deviations of various items from the median or mean of that series.

Median is preferred since the sum of the diviations from the median is less than that from the mean. So the values of mean deviation calculated from median is usually less than that calculated from mean. Mode is not considered, as its value is indeterminate.

Mean deviation is known as *First Moment* of dispersion.

Computations of Mean Deviation.

FOR INDIVIDUAL OBSERVATION : The formula is as follows :

$$\text{Mean Deviation (M.D.)} = \frac{\Sigma |D|}{n},$$

where $|D|$ within two vertical lines denotes deviations from mean (or median), ignoring algebraic signs (*i.e.*, + and −).

Steps to Find M.D. :

(1) Find mean or median ;

(2) Take deviations ignoring ± signs ;

(3) Get total of deviations ;

(4) Divide the above total by the number of items.

Example.

To find the mean deviation of the following data about mean and median (Rs.) 2, 6, 11, 14, 16, 19, 23.

Computation of Mean Deviation

About Mean			About Median		
Serial no.	x (Rs.)	Dev. from A.M. ignoring ± signs $\|D\|$	Serial no.	x (Rs.)	Dev. form Med. ignoring ± signs $\|D\|$
1	2	11	1	2	12
2	6	7	2	6	8
3	11	2	3	11	3
4	14	1	4	14	0
5	16	3	5	16	2
6	19	6	6	19	5
7	23	10	7	23	9
Total	—	40	Total	—	39

$$\text{A. M.} = \tfrac{1}{7}(2+6+11+14+16+19+23) = \tfrac{1}{7} \times 91 = \text{Rs. } 13.$$

$$\text{Median} = \text{size of } \frac{7+1}{2}\text{th item} = \text{size of 4th item} = \text{Rs. } 14.$$

$$\text{Mean deviation (about mean)} = \frac{\Sigma|D|}{n} = \frac{40}{7} = \text{Rs. } 5{\cdot}71.$$

$$\text{Mean deviation (about median)} = \frac{\Sigma|D|}{n} = \frac{39}{7} = \text{Rs. } 5{\cdot}57.$$

Note. The sum of deviation ($\Sigma|D|$) about median is 39, less than $|D|$ about mean (= 40). Also M.D. about median (*i.e.*, 5·57) is less than that about mean (*i.e.*, 5·71).

Coefficient of Mean Deviation.

$$\text{About mean, Coefficient of M. D.} = \frac{\text{M. D.}}{\text{mean}} = \frac{5{\cdot}71}{13} = {\cdot}44 \text{ (app.)}$$

$$\text{About median, Coefficient of M. D.} = \frac{\text{M. D.}}{\text{median}} = \frac{5{\cdot}57}{14} = {\cdot}40 \text{ (app.)}$$

For Discrete Series :

The formula for computing M. D. is

$$\text{M. D.} = \frac{\Sigma f|\text{D}|}{\Sigma f},$$

where $|\text{D}|$ = deviations from mean (or median) ignoring ± signs.

About Mean.

Example : To calculate mean deviation of the following series :

(Marks) x :	5	10	15	20	25	Total
(Student) f :	6	7	8	11	8	40

—Find also the coefficient of dispersion.

Computation of Mean Diviation (About Mean).

Marks		Dev. from assumed mean (15)	Step deviation		Deviation from actual mean (16)	
x	f	d	$d' = d/5$	fd'	$\|\text{D}\|$	$f\|\text{D}\|$
(1)	(2)	(3)	(4)	(5) = (2) × (4)	(6)	(7) = (2) × (6)
5	6	−10	−2	−12	11	66
10	7	−5	−1	−7	6	42
15	8	0	0	0	1	8
20	11	5	1	11	4	44
25	8	10	2	16	9	72
Total	40	—	—	8	—	232

$$\text{A. M.} = \text{A} + \frac{\Sigma fd'}{\Sigma f} \times i = 15 + \frac{8}{40} \times 5 = 15 + 1 = 16 \text{ marks.}$$

$$\text{M.D.} = \frac{\Sigma f|\text{D}|}{\Sigma f} = \frac{232}{40} = 5{\cdot}8 \text{ marks.}$$

$$\text{Coefficient of dispersion (about mean)} = \frac{\text{M.D.}}{\text{mean}} = \frac{5{\cdot}8}{16} = {\cdot}363.$$

About Median.

Example. The same example as given above.

Computation of Mean Deviation (About Median)

Marks		Cum. freq.	Dev. from median (15)	
x	f	$c.f$	$\lvert D \rvert$	$f\lvert D \rvert$
5	6	6	10	60
10	7	13	5	35
15	8	21	0	0
20	11	32	5	55
25	8	40 (=N)	10	80
Total	40	—	—	230

Median = size of the $\frac{40+1}{2}$th item = size of 20·5th item = 15 marks

$$\text{M. D.} = \frac{\Sigma f \lvert D \rvert}{\Sigma f} = \frac{230}{40} = 5{\cdot}75 \text{ marks.}$$

$$\text{Coefficient of dispersion (about median)} = \frac{\text{M. D.}}{\text{median}} = \frac{5{\cdot}75}{15} = {\cdot}383.$$

FOR CONTINUOUS SERIES :

Calculations is similar to the above process. The only difference is that we are to take the deviations (in case of M. D. about median) from the middle points of the various class-intervals.

About Mean.

Example. Measurements of the lengths in feet of 50 iron rods are distributed as follows :

Class-boundary	*Frequency*
2·35—2·45	1
2·45—2·55	4
2·55—2·65	7
2·65—2·75	15
2·75—2·85	11
2·85—2·95	10
2·95—3·05	2

Find, to two decimal places, the value of the mean deviation.

[C. U. M. Com. 1965]

Computation of Mean Deviation (About Mean)

Class-boundary (*ft*)	*f*	Mid-value *x*	Dev. from ass. mean 2·70 *d*	$d'=d/{\cdot}1$	fd'	Lev. from mean 2·738 |D|	*f*|D|
(1)	(2)	(3)	(3)	(4)	(5)=(2)×(4)	(6)	(7)=(2)×(6)
2·35—2·45	1	2·40	−·30	−3	−3	·338	·338
2·45—2·55	4	2·50	−·20	−2	−8	·238	·952
2·55—2·65	7	2·60	−·10	−1	−7	·138	·966
2·65—2·75	15	2·70	0	0	0	·038	·570
2·75—2·85	11	2·80	·10	1	11	·062	·682
2·85—2·95	10	2·90	·20	2	20	·162	1·620
2·95—3·05	2	3·00	·30	3	6	·262	·524
Total	50	—	—	—	19	—	5·652

$$\text{A.M.}=\text{A}+\frac{\Sigma fd'}{\Sigma f}\times i=2{\cdot}70+\frac{19}{50}\times({\cdot}1)=2{\cdot}70+{\cdot}038=2{\cdot}738\text{ ft.}$$

$$\text{M.D. (About Mean)}=\frac{\Sigma f|\text{D}|}{\Sigma f}=\frac{5{\cdot}652}{50}={\cdot}11304={\cdot}11\text{ ft. (2 dec. pl.)}$$

About Median.

Example. Calculate mean deviation from median and the corresponding coefficient of dispersion for the following data :

Height in inches	*No. of saplings*	*Height in inches*	*No. of saplings*
4·5 and under 5·5	2	8·5 and under 9·5	25
5·5 " " 6·5	6	9·5 " " 10·5	20
6·5 " " 7·5	12	10·5 " " 11·5	10
7·5 " " 8·5	18	11·5 " " 12·5	7

Take $\frac{N+1}{2}$ as the rank of median. [C. A. 1963]

Computation of Mean Deviation (About Median)

Height (inches)	f	Cum. freq. $c.f$	Mid-value x	Dev. from median (9) $\lvert D \rvert$	$f\lvert D \rvert$
(1)	(2)	(3)	(4)	(5)	(6)=(2)×(5)
4·5— 5·5	2	2	5	4	8
5·5— 6·5	6	8	6	3	18
6·5— 7·5	12	20	7	2	24
7·5— 8·5	18	38	8	1	18
8·5— 9·5	25	63	9	0	0
9·5—10·5	20	83	10	1	20
10·5—11·5	10	93	11	2	20
11·5—12·5	7	100 (=N)	12	3	21
Total	100	—	—	—	129

$$\text{Median} = \text{size of } \frac{N+1}{2}\text{th item} = \text{size of } \frac{101}{2}\text{th item}$$

$$= \text{size of } 50{\cdot}5\text{th item.}$$

So median lies in the class (8·5—9·5).

$$\therefore \quad \text{Median} = 8{\cdot}5 + \frac{9{\cdot}5 - 8{\cdot}5}{25}(50{\cdot}5 - 38) = 8{\cdot}5 + {\cdot}5 = 9 \text{ inches.}$$

$$\therefore \quad \text{M. D.} = \frac{\Sigma f|D|}{\Sigma f} = \frac{129}{100} = 1{\cdot}29 \text{ inches.}$$

$$\text{Coefficient of dispersion (about median)} = \frac{\text{M. D.}}{\text{median}} = \frac{1{\cdot}29}{9} = {\cdot}143.$$

Note : In continuous series, $\frac{N}{2}$ is used as rank of median. But here we are asked to use $\frac{N+1}{2}$.

Computation of Mean Deviation—Short-cut Method.

When median (or mean) are in fraction, calculation becomes difficult. In such a case, following short-cut method is used for computation.

$$\text{M. D.} = \frac{\Sigma u - \Sigma l}{n}, \text{ where } u = \text{items greater than median}$$
$$l = \text{ ,, lower ,, ,,}$$

Example : To calculate the mean deviation about median in the following series of marks : 15, 14, 17, 20, 12, 24, 21, 27, 26, 30.

Arrangement : 12, 14, 15, 17, 20, 21, 24, 26, 27, 30, $n = 10$

$$\text{Median} = \text{size of } \frac{n+1}{2}\text{ th item} = \text{size of } 5{\cdot}5\text{th item} = 20{\cdot}5 \text{ marks.}$$

Items greater than median (20·5) are 21, 24, 26, 27, 30, their total = 128. Items less than median are 12, 14, 15, 17, 20, their total = 78.

$$\therefore \text{ M. D.} = \frac{128 - 78}{10} = \frac{50}{10} = 5 \text{ marks.}$$

Advantages and Disadvantages of Mean Deviation.

Advantages :

(1) It is based on all the observations. Any change in any item would change the value of mean deviation.

(2) It is readily understood. It is the average of the deviations from a measure of central tendency.

(3) Mean deviation is less affected by the extreme items than the standard deviation.

(4) It is simple to understand and easy to compute.

Disadvantages :

(1) Mean deviation ignores the algebraic signs of the deviations, and as such it is not capable of further algebraic treatment.

(2) It is not an accurate measure, particularly when it is calculated from mode.

(3) It is not popular as standard deviations.

Uses of Mean Deviation.

Because of simplicity in computation, it has drawn the alteration of economists and businessmen. It is useful in reports meant for public.

Standard Deviation.

In calculating mean deviation we ignored the algebraic signs, which is mathematically illogical. This drawback is removed in calculating standard deviation, usually denoted by 'σ' (read as sigma).

Definition :

Standard deviation is the square root of the arithmetic average of the squares of all the deviations from the mean. In short, it may be defined as the root-mean-square deviation from the mean.

If $\bar{x}$ is the mean of $x_1, x_2, \ldots\ldots, x_n$, then σ is defined by

$$\sqrt{\left[\frac{1}{n}\left\{(x_1-\bar{x})^2+\cdots\cdots+(x_n-\bar{x})^2\right\}\right]}=\sqrt{\frac{1}{n}\left(n\bar{x}^2-2\bar{x}\,\Sigma x+\Sigma x^2\right)}$$

$$=\sqrt{\frac{1}{n}\left(n\bar{x}^2-2\bar{x}.n\bar{x}+\Sigma x^2\right)}=\sqrt{\frac{1}{n}\left(\Sigma x^2-n\bar{x}^2\right)}=\sqrt{\left\{\frac{\Sigma x^2}{n}-\left(\frac{\Sigma x}{n}\right)^2\right\}}$$

Computation of Standard Deviation.

For Individual Observations :

Computation may be done in two ways—(*a*) by taking deviations from actual mean, (*b*) by taking deviation from assumed mean.

(*a*) Steps to follow—(1) Find the actual mean,

(2) Find the deviations from the mean,

(3) Make squares of the deviations and add up,

(4) Divide the addition by total number of items and find square root.

(*b*) Steps to follow—(1) Find the deviations of the items from an assumed mean and denote it by *d*. Find also Σd,

(2) Square the deviations, find Σd^2,

(3) Apply the following formula to find standard deviations,

$$\text{S. D. } (\sigma)=\sqrt{\frac{\Sigma d^2}{n}-\left(\frac{\Sigma d}{n}\right)^2}.$$

Example**:** The table below shows the marks obtained by 10 students in a certain test. Calculate the standard deviation by both the above methods.

Roll No. :	1	2	3	4	5	6	7	8	9	10
Marks :	43	48	65	57	31	60	37	48	78	58

Computation of Standard Deviation

	Method (*a*)			Method (*b*)		
Roll No.	Marks	Dev. from actual mean	Square of dev.	Marks	Dev. from assumed mean	Square of dev.
		52·5 (*d*)	(d^2)		50 (*d*)	(d^2)
1	43	−9·5	90·25	43	−7	49
2	48	−4·5	20·25	48	−2	4
3	65	12·5	156·25	65	15	225
4	57	4·5	20·25	57	7	49
5	31	−21·5	462·25	31	−19	361
6	60	7·5	56·25	60	10	100
7	37	−15·5	240·25	37	−13	169
8	48	−4·5	20·25	48	−2	4
9	78	25·5	650·25	78	28	784
10	58	5·5	30·25	58	8	64
Total	525		1746·50	525	$\Sigma d=25$	$\Sigma d^2=1809$

For Method (*a*),

$$\text{A. M.} = \frac{1}{10} \times 525 = 52{\cdot}5 \text{ marks.}$$

$$\text{S. D. } (\sigma) = \sqrt{\frac{1746{\cdot}50}{10}} = \sqrt{174{\cdot}65} = 13{\cdot}21 \text{ marks.}$$

Here the average marks are 52·5, and they deviate on an average from the average by 13·21 marks.

For Method (*b*),

$$\text{S. D. } (\sigma) = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2} = \sqrt{\frac{1809}{10} - \left(\frac{25}{10}\right)^2} = \sqrt{180{\cdot}9 - 6{\cdot}25}$$

$$= \sqrt{174{\cdot}65} = 13{\cdot}21 \text{ marks.}$$

Note. If the actual mean is in fraction, then it is better to take deviations from an assumed mean, for avoiding too much calculations.

For Discrete Series :

There are three methods, given below, for computing Standard Deviation :

(*a*) Actual Mean, (*b*) Assumed Mean, (*c*) Step Deviation.

For (*a*), the following formula is used :

This method is used rarely because if the actual mean is in fractions, calculations take much time.

$$\sigma = \sqrt{\frac{\Sigma fx^2}{\Sigma f}}, \text{ where } x = (\text{X} - \overline{\text{X}}).$$

For (*b*), the followings are the *steps* to be used :

(i) Find the deviations (from ass. mean), denote it by d,

(ii) Obtain Σfd,

(iii) Find d^2 and then Σfd^2,

then use the formula :

$$\sigma = \sqrt{\frac{\Sigma fd^2}{\Sigma f} - \left(\frac{\Sigma fd}{\Sigma f}\right)^2}$$

Example. Find the standard deviation of the following series :

x :	10	11	12	13	14	Total
f :	3	12	18	12	3	48

[C. A. May 1963]

Calculation of Standard Deviations

x	f	Dev. from ass. mean (12) d	fd	d^2	fd^2
(1)	(2)	(3)	(4) = (2) × (3)	(5) = (3) × (3)	(6) = (2) × (5)
10	3	−2	−6	4	12
11	12	−1	−12	1	12
12	18	0	0	0	0
13	12	1	12	1	12
14	3	2	6	4	12
	48	–	0	–	48

$$\sigma = \sqrt{\left\{\frac{\Sigma fd^2}{\Sigma f} - \left(\frac{\Sigma fd}{\Sigma f}\right)^2\right\}} = \sqrt{\left(\frac{48}{48} - \frac{0}{48}\right)} = \sqrt{1} = 1.$$

For (*c*), The following formula is used :

The idea will be clear from example shown below :

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{\Sigma f} - \left(\frac{\Sigma fd'}{\Sigma f}\right)^2} \times i, \quad \text{where } d' = \text{step deviation,}$$
$$i = \text{common factor.}$$

Example :

Find the standard deviation for the following distribution :

x :	4·5	14·5	24·5	34·5	44·5	54·5	64·5
f :	2	3	5	17	12	7	4

$$\sigma = \sqrt{\left\{\frac{\Sigma fd'^2}{\Sigma f} - \left(\frac{\Sigma fd'}{\Sigma f}\right)^2\right\}} \times i = \sqrt{\left\{\frac{111}{50} - \left(\frac{21}{50}\right)^2\right\}} \times 10$$

(Computation Table is shown in the next page)

$$= \sqrt{(2·22 - ·1764)} \times 10 = 1·4295 \times 10 = 14·295.$$

Calculation of Standard Deviation

x	f	d	$d' = d/10$	fd'	fd'^2
4·5	2	−30	−3	−6	18
14·5	3	−20	−2	−6	12
24·5	5	−10	−1	−5	5
34·5	17	0	0	0	0
44·5	12	10	1	12	12
54·5	7	20	2	14	28
64·5	4	30	3	12	36
	$\Sigma f = 50$	—	—	$\Sigma fd' = 21$	$\Sigma fd'^2 = 111$

For Continuous Series :

Any method discussed above (for discrete series) can be used in this case. Of course, step deviation method is convenient to use. From the following example, procedure of calculation will be clear.

Example : Find the standard deviation from the following frequency distribution :

Weight (lb.) :	131—140	141—150	151—160	161—170	171—180	181—190	191—200	211—240
No. of persons :	2	5	4	9	7	5	3	1

[I.C.W.A. Jan. 1971]

Calculation of Standard Deviation

Weight (*lb.*)	f	Mid-value x	d	$d' = d/5$	fd'	fd'^2
131—140	2	135·5	−30	−6	−12	72
141—150	5	145·5	−20	−4	−20	80
151—160	4	155·5	−10	−2	−8	16
161—170	9	165·5	0	0	0	0
171—180	7	175·5	10	2	14	28
181—190	5	185·5	20	4	20	80
191—210	3	200·5	35	7	21	147
211—240	1	225·5	60	12	12	144
Total	36	—	—	—	27	567

$$\sigma = \sqrt{\left\{\frac{\Sigma fd'^2}{\Sigma f} - \left(\frac{\Sigma fd'}{\Sigma f}\right)^2\right\}} \times i = \sqrt{\left\{\frac{567}{36} - \left(\frac{27}{36}\right)^2\right\}} \times 5$$

$$= \sqrt{(15{\cdot}75 - {\cdot}5626)} \times 5 = \sqrt{15{\cdot}1874} \times 5$$

$$= 3{\cdot}897 \times 5 = 19{\cdot}485 = 19{\cdot}50 \text{ lbs. (app.)} \quad \text{(Calculation by log table)}$$

Note. If we are to find the mean, then

$$\text{A.M.} = \text{A} + \frac{\Sigma fd'}{\Sigma f} \times i = 165{\cdot}5 + \frac{27}{36} \times 5 = 165{\cdot}5 + {\cdot}75 \times 5$$

$$= 165{\cdot}5 + 3{\cdot}75 = 169{\cdot}25 \text{ lbs.}$$

Mathematical Properties of Standard Deviation.

(1) COMBINED STANDARD DEVIATION :

We can also calculate the combined standard deviation for two or more groups, similar to mean of composite group. The required formula is as follows :

$$\sigma_{12} = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_1d_1^2 + n_2d_2^2}{n_1 + n_2}}.$$

Where, σ_{12} = combined standard deviation of two groups,

σ_1 = standard deviation of 1st group,

σ_2 = standard deviation of 2nd group,

$d_1 = \bar{x}_1 - \bar{x}_{12}$; $d_2 = \bar{x}_2 - \bar{x}_{12}$

For Three Groups,

$$\sigma_{123} = \sqrt{\left\{\frac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_3\sigma_3^2 + n_1d_1^2 + n_2d_2^2 + n_3d_3^2}{n_1 + n_2 + n_3}\right\}}$$

where, $d_1 = \bar{x}_1 - \bar{x}_{123}$; $d_2 = \bar{x}_2 - \bar{x}_{123}$; $d_3 = \bar{x}_3 - \bar{x}_{123}$.

Example **:** Two samples of sizes 40 and 50 respectively have the same mean 53, but different standard deviations 19 and 8 respectively. Find the standard deviations of the combined sample of size 90.

[C. A. Nov. 1963]

Here, $n_1 = 40$, $\bar{x}_1 = 53$, $\sigma_1 = 19$

$n_2 = 50$, $\bar{x}_2 = 53$, $\sigma_2 = 8$.

Now, $$\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{40 \times 53 + 50 \times 53}{40 + 50}$$

$$= \frac{2120 + 2650}{90} = \frac{4770}{90} = 53.$$

Now, $d_1 = \bar{x}_1 - \bar{x}_{12} = 53 - 53 = 0$, $d_2 = 0$,

$$\therefore \quad \sigma_{12} = \sqrt{\left\{\frac{40(19)^2 + 50(8)^2 + 40(0)^2 + 50(0)^2}{40 + 50}\right\}}$$

$$= \sqrt{\left(\frac{14440 + 3200}{90}\right)} = \sqrt{\frac{17640}{90}} = \sqrt{196} = 14.$$

Example **:** The number of workers employed, the Mean Wages (in Rs.) per month and the Standard Deviation (in Rs.) in each section of a factory are given below. Calculate the Mean Wages and Standard Deviation of all the workers taken together.

Section	*No. of workers employed*	*Mean wages (in Rs.)*	*Standard deviation (in Rs.)*
A	50	113	6
B	60	120	7
C	90	115	8

[I.C.W.A. Jan. 1964]

We know $$\bar{x}_{123} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{n_1 + n_2 + n_3} = \frac{50 \times 113 + 60 \times 120 + 90 \times 115}{50 + 60 + 90}$$

$$= \frac{5650 + 7200 + 10350}{200} = \frac{23200}{200} = \text{Rs. } 116.$$

Now, $d_1 = \bar{x}_1 - \bar{x}_{123} = 113 - 116 = -3,$
$d_2 = \bar{x}_2 - \bar{x}_{123} = 120 - 116 = 4,$
$d_3 = \bar{x}_3 - \bar{x}_{123} = 115 - 116 = -1.$

$$\sigma_{123} = \sqrt{\left\{\frac{50(6)^2 + 60(7)^2 + 90(8)^2 + 50(-3)^2 + 60(4)^2 + 90(-1)^2}{50 + 60 + 90}\right\}}$$

$$= \sqrt{\left\{\frac{1800 + 2940 + 5760 + 450 + 960 + 90}{200}\right\}} = \sqrt{\frac{12000}{200}}$$

$$= \sqrt{60} = \text{Rs. } 7{\cdot}75.$$

(2) **Prove that the Standard Deviation does not depend on the choice of origin.**

For the n observations $x_1, x_2, \ldots, x_n$ let $d_1, d_2, \ldots, d_n$ are respective quantities obtained by shifting the origin to any arbitrary constant, say A, so that $d_i = x_i - A$ (for $i = 1, 2, \ldots, n$). Now we are to show $\sigma_x = \sigma_d$.

We know, $\sigma_x^2 = \Sigma(x_i - \bar{x})^2/n$, where $\bar{x} = \Sigma x_i/n$

Now, $d_i = x_i - A$ so that $\Sigma d_i = \Sigma x_i - \Sigma A$ (taking Σ to both sides).

Again $\dfrac{\Sigma d_i}{n} = \dfrac{\Sigma x_i}{n} - \dfrac{\Sigma A}{n}$ (dividing by n)

or, $\bar{d} = \bar{x} - A$ or, $\bar{x} = A + \bar{d}$

Now, $x_i - \bar{x} = (A + d_i) - (A + \bar{d}) = d_i - \bar{d}$

So $\sigma_x^2 = \Sigma(x_i - \bar{x})^2/n = \Sigma(d_i - \bar{d})^2/n = \sigma_d^2$

$\therefore \quad \sigma_x = \sigma_d.$

(3) **Prove that the Standard Deviation calculated from two values x_1 and x_2 of a variable x is equal to half their difference.**

[I. C. W. A. Jan. 1968]

We know $\sigma^2 = \dfrac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2}{2}$, according to definition of σ, and where $\bar{x} = \frac{1}{2}(x_1 + x_2)$ *i.e.*, $\bar{x}$ is A.M. of x_1 and x_2

$$\text{or} \quad \sigma^2 = \tfrac{1}{2}\left[\left(x_1 - \frac{x_1 + x_2}{2}\right)^2 + \left(x_2 - \frac{x_1 + x_2}{2}\right)^2\right]$$

(putting the value of $\bar{x}$)

$$= \tfrac{1}{2}\left[\left(\frac{x_1 - x_2}{2}\right)^2 + \left(\frac{x_2 - x_1}{2}\right)^2\right]$$

$$= \tfrac{1}{2}[\tfrac{1}{4}\{(x_1 - x_2)^2 + (x_1 - x_2)^2\}] = \tfrac{1}{4}(x_1 - x_2)^2$$

$\therefore \quad \sigma = \frac{1}{2}(x_1 - x_2)$, since σ is always positive.

(4) Show that if $\bar{x}$ is the Arithmetic Mean of the quantities $x_1, x_2, \ldots, x_n$ then $\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$.

[C. U. M. Com. 1963 ; I.C.W.A. Jan. 1969]

$$\sum_1^n(x_i-\bar{x})^2 = \sum_1^n(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_1^n x_i^2 - 2\bar{x}\sum_1^n x_i + \sum_1^n \bar{x}^2$$

$$= \sum_1^n x_i^2 - 2\bar{x}\, n\bar{x} + n\bar{x}^2 \quad [\bar{x} = \Sigma x_i/n \text{ or, } \Sigma x_i = n\bar{x}]$$

$$= \sum_1^n x_i^2 - 2n\bar{x}^2 + n\bar{x}^2$$

$$= \sum_1^n x_i^2 - n\bar{x}^2.$$

(5) Prove that the standard deviation is independent of any change of origin, but is dependent on the change of scale.

[I. C. W. A. Jan. 1971 ; Dec. 78]

Without Frequency :

For the n observations $x_1, x_2, \ldots, x_n$, let the origin be changed to A and the scale to d, then $y_i = \frac{x_i - A}{d}$ or, $x_i = A + dy_i$, which means $y_1, y_2, \ldots, y_n$ are the deviations of $x_1, x_2, \ldots, x_n$ from an arbitrary constant A, in units of another constant d.

Now, $\bar{x} = A + d\bar{y}$ *i.e.*, mean of x's $= A + d$ (mean of y's)

Again, $x_i - \bar{x} = (A + dy_i) - (A + d\bar{y}) = d(y_i - \bar{y})$

$$\sigma_x^2 = \frac{\Sigma(x_i-\bar{x})^2}{n} = \frac{\Sigma\{d(y_i-\bar{y})\}^2}{n} = \frac{d^2\Sigma(y_i-\bar{y})^2}{n} = d^2\sigma_y^2$$

$\therefore$ $\sigma_x = d\sigma_y$ (A is absent, but d is present).

This shows S. D. is uneffected by any change of origin, but depends on scale.

With Frequency :

S. D. $(\sigma_x) = \sqrt{\frac{\Sigma f(x-\bar{x})^2}{\Sigma f}}$, $\bar{x}$ = actual A. M. (weighted mean) of variates x. Changing the origin to A(say), let $u = x - A$ or, $x = u + A$ *i.e.*, $\bar{x} = \bar{u} + A$.

Now, $\sigma_x = \sqrt{\frac{\Sigma f(u + A - \bar{u} - A)^2}{\Sigma f}} = \sqrt{\frac{\Sigma f(u-\bar{u})^2}{\Sigma f}} = \sigma_u$

$\therefore$ $\sigma_x = \sigma_u$ *i.e.*, S. D. is uneffected by change of origin.

Now for change of scale, let $u=\frac{x-A}{d}$, d = width of class

$$\text{or, } ud=x-A$$
$$\text{or, } x=ud+A \text{ i.e., } \bar{x}=\bar{u}d+A.$$

Substituting the value in σ_x we find,

$$\sigma_x=\sqrt{\frac{\Sigma f(x-\bar{x})^2}{\Sigma f}}=\sqrt{\frac{\Sigma f(ud+A-\bar{u}d-A)^2}{\Sigma f}}=\sqrt{\frac{\Sigma f(ud-\bar{u}d)^2}{\Sigma f}}$$
$$=\sqrt{\frac{\Sigma f(u-\bar{u})^2d^2}{\Sigma f}}=d\times\sqrt{\frac{\Sigma f(u-\bar{u})^2}{\Sigma f}}=d\sigma_u$$

i.e., S. D. is affected by change of scale.

(6) Find the A.M. and Standard Deviation of the first n natural numbers.

1, 2, 3, ..., n are the first n natural numbers.

Now, A.M. $=\frac{1+2+3+\cdots+n}{n}=\frac{n(n+1)/2}{n}$ (for the sum formula, see A.P. chapter)

$$=\frac{n(n+1)}{2n}=\frac{n+1}{2} \quad \cdots \quad (1)$$

Again we know, $1^2+2^2+3^2+\cdots+n^2=\frac{n(n+1)(2n+1)}{6}$ ... (2)

(see A. P. chapter)

Now, $\sigma=\sqrt{\left\{\frac{\Sigma x^2}{n}-\left(\frac{\Sigma x}{n}\right)^2\right\}}$ (see the formula given in def. of S. D.)

i.e., $\sigma^2=\frac{\Sigma x^2}{n}-\left(\frac{\Sigma x}{n}\right)^2=\frac{n(n+1)(2n+1)}{6n}-\left(\frac{n+1}{2}\right)^2$ by (1), (2)

$$=\frac{(n+1)(2n+1)}{6}-\frac{(n+1)^2}{4}=\frac{2(n+1)(2n+1)-3(n+1)^2}{12}$$
$$=\frac{(n+1)(4n+2-3n-3)}{12}=\frac{(n+1)(n-1)}{12}=\frac{n^2-1}{12}$$

$\therefore \quad \sigma\ (s.\ d.)=\sqrt{\frac{n^2-1}{12}}.$

Example**:** Find A. M. and S. D. of the natural numbers 1 to 11.

A. M. $=\frac{11+1}{2}$ (as $n=11$) $=\frac{12}{2}=6$

S. D. $(\sigma)=\sqrt{\frac{11^2-1}{12}}=\sqrt{\frac{121-1}{12}}=\sqrt{\frac{120}{12}}=\sqrt{10}=3{\cdot}1623.$

Charlier's Check of Accuracy.

Checking accuracy of computation can also be applied in case of standard deviation, by applying the following equation :

$$\Sigma[f(d'+1)^2] = \Sigma(fd'^2) + 2\Sigma(fd') + \Sigma f$$

Example **:** Apply Charlier's Check of Accuracy and hence find Standard Deviation and Arithmetic Mean.

Wages (Rs.) :	0—10	10—20	20—30	30—40	40—50
No. of Workers :	2	6	20	14	8

Calculation of Standard Deviation and Mean

Wages (Rs.)	f	x	d	d'	fd'	fd'^2	$f(d'+1)^2$
0—10	2	5	−20	−2	−4	8	2
10—20	6	15	−10	−1	−6	6	0
20—30	20	25	0	0	0	0	20
30—40	14	35	10	1	14	14	56
40—50	8	45	20	2	16	32	72
Total	50	—	—	—	20	60	150

Now, $\Sigma[f(d'+1)^2] = 150$

Again, $\Sigma(fd'^2) + 2\Sigma(fd') + \Sigma f = 60 + 2 \times 20 + 50 = 60 + 40 + 50 = 150.$

Hence the calculations are correct.

Now, $\sigma = \sqrt{\frac{\Sigma fd'^2}{\Sigma f} - \left(\frac{\Sigma fd'}{\Sigma f}\right)^2} \times i = \sqrt{\left\{\frac{60}{50} - \left(\frac{20}{50}\right)^2\right\}} \times 10$

$= 5{\cdot}099 \times 10 = \text{Rs. } 50{\cdot}99$

$\text{A.M.} = A + \frac{\Sigma fd'}{\Sigma f} \times i = 25 + \frac{20}{50} \times 10 = 25 + 4 = \text{Rs. } 29.$

Variance.

The square of the Standard Deviation is known as *Variance.*

Coefficient of Variation.

It is the ratio of the Standard Deviation to the Mean expressed as percentage. This relative measure was first suggested by Professor Karl Pearson. According to him, coefficient of variation is the percentage variation in the Mean, while Standard Deviation is the total variation in the Mean.

Symbolically, Coefficient of variation $(V)=\frac{\sigma}{\bar{x}}\times 100=$ Coefficient of standard deviation × 100.

Note. The coefficient of variation is also known as coefficient at variability.

***Example*:**

If Mean and Standard Deviations of a series are respectively 40 and 10, then the coefficient of variations would be $\frac{10}{40}\times 100=25\%$, which means the Standard Deviation is 25% of the mean.

Note. For comparing variability of two or more series, it is used commonly. A series of having coefficient of variation greater, is said to be more variable, *i.e.*, less uniform, less stable or less consistent. Again a series, having coefficient of variation lesser is said to be less variable, *i.e.*, more uniform, more stable or more consistent.

***Example*:**

From the marks given below obtained by two students taking the same course, find out who is more intelligent and who is more consistent student.

A :	58	59	60	65	66	52	75	31	46	48
B :	56	87	89	46	93	65	44	54	78	68

In order to find out the more consistent between the two students, we are to compute the mean marks and then the coefficient of variations for comparison.

For Student A : $\bar{x}=\frac{\Sigma x}{n}=\frac{560}{10}=56$ marks.

Now, $\sigma=\sqrt{\frac{\Sigma d^2}{n}}=\sqrt{\frac{1376}{10}}=\sqrt{137\cdot6}=11\cdot73$ marks.

(Computation Table is shown in the next page).

Computation of Mean and Standard Deviation

For Student A			*For Student* B		
Marks x	d	d^2	Marks x	d	d^2
58	2	4	56	-12	144
59	3	9	87	19	361
60	4	16	89	21	441
65	9	81	46	-22	484
66	10	100	93	25	625
52	-4	16	65	-3	9
75	19	361	44	-24	576
31	-25	625	54	-14	196
46	-10	100	78	10	100
48	-8	64	68	0	0
560		1376	680		2936

Again V (coefficient of variation) $= \frac{11{\cdot}73}{56} \times 100 = 20{\cdot}94\%$.

For Student B : $\bar{x} = \frac{\Sigma x}{n} = \frac{680}{10} = 68$ marks.

Now, $\sigma = \sqrt{\frac{\Sigma d^2}{n}} = \sqrt{\frac{2936}{10}} = \sqrt{293{\cdot}6} = 17{\cdot}14$ marks.

Again $V = \frac{17{\cdot}14}{68} \times 100 = 25{\cdot}21\%$.

Average marks obtained by Student B are higher than that of Student A, so Student B is more intelligent. Again since coefficient of variation of Student A is less than that of Student B, so Student A is more consistent.

***Example*:** Suppose that samples of polythene bags from two manufactures, A and B are tested by a prospective buyer for bursting pressure, with the following results :

Bursting pressure (16)	*Numbers of bags*	
	A	B
5·0— 9·9	2	9
10·0—14·9	9	11
15·0—19·9	29	18
20·0—24·9	54	32
25·0—29·9	11	27
30·0—34·9	5	13
	110	110

Which set of bags has the highest average bursting pressure ? Which has more uniform pressure ? If prices are the same, which manufacturer's bags would be preferred by the buyer ? Why ?

[B. Com. Delhi 1966]

Computation of Mean and Standard Deviation

Bursting Pressure (16)	*Mid-pt.* x	$d'=d/5$	*For bag A*			*For bag B*		
			f	fd'	fd'^2	f	fd'	fd'^2
4·95— 9·95	7·45	−2	2	−4	8	9	−18	36
9·95—14·95	12·45	−1	9	−9	9	11	−11	11
14·95—19 95	17·45	0	29	0	0	18	0	0
19·95—24·95	22·45	1	54	54	54	32	32	32
24·95—29·95	27·45	2	11	22	44	27	54	108
29·95—34·95	32·45	3	5	15	45	13	39	117
Total	—	—	110	78	160	110	96	304

For bag A : $\text{Mean} = A + \dfrac{\Sigma fd'}{\Sigma f} \times i = 17{\cdot}45 + \dfrac{78}{110} \times 5 = 17{\cdot}45 + 3{\cdot}55 = 21 \text{ lb.}$

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{\Sigma f} - \left(\frac{\Sigma fd'}{\Sigma f}\right)^2} \times i = \sqrt{\frac{160}{110} - \left(\frac{78}{110}\right)^2} \times 5 = \sqrt{(1{\cdot}455 - {\cdot}503)} \times 5$$

$$= \sqrt{{\cdot}952} \times 5 = {\cdot}976 \times 5 = 4{\cdot}880 \text{ lb.}$$

$$V = \frac{\sigma}{\text{Mean}} \times 100 = \frac{4{\cdot}880}{21} \times 100 = 23{\cdot}24\%.$$

For bag B : $\text{Mean} = 17{\cdot}45 + \dfrac{96}{110} \times 5 = 17{\cdot}45 + 4{\cdot}36 = 21{\cdot}81 \text{ lb.}$

$$\sigma = \sqrt{\left\{\frac{304}{110} - \left(\frac{96}{110}\right)^2\right\}} \times 5 = \sqrt{(2{\cdot}764 - {\cdot}762)} \times 5 = \sqrt{2{\cdot}002} \times 5 = 1{\cdot}417 \times 5$$

$$= 7{\cdot}085 \text{ lb.}$$

$$V = \frac{7{\cdot}085}{21{\cdot}81} \times 100 = 32{\cdot}48\%.$$

The bags of Manufacturer B have the highest bursting pressure, which is clear from the averages calculated above. Again the bags of Manufacturer A have more uniform pressure, since the coefficient of variation is less than that of Manufacturer B. If again, the prices are same, the bags of Manufacturer A should be preferred by the buyer because they have more uniform pressure.

Example **:** An analysis of the monthly wages paid to workers in two firms, A and B, belonging to the same industry gives the following results :

	Firm A	*Firm* B
No. of wage-earners	586	648
Average monthly wages	Rs. 52·5	Rs. 47·5
Variance of the distribution of wages	100	121

(*a*) Which firm A or B pays out the largest amount as monthly wages ?

(*b*) Which firm A or B has greater variability in individual wages ?

(*c*) Find the average monthly wages and the standard deviation of the wages of all the workers in two firms A and B taken together.

[C. U. B.A. (Econ.) 1970 ; Madras B. Com. 1962 ; I.C.W.A. Jan. 1965]

(*a*) *For firm* A : Total wages $= 586 \times 52{\cdot}5 =$ Rs. 30,765.

For firm B : Total wages $= 648 \times 47{\cdot}5 =$ Rs. 30,780.

∴ Firm B pays largest amount.

(*b*) *For firm* A : $\sigma^2 = 100$. ∴ $\sigma =$ Rs. 10.

$$\text{Now,}\quad V=\frac{\sigma}{\text{Mean}}\times 100=\frac{10}{52\text{·}5}\times 100=19\text{·}04.$$

$$\textit{For Firm B}:\quad V=\frac{11}{47\text{·}5}\times 100=23\text{·}16\quad(\text{here }\sigma=\text{Rs. }11).$$

$\therefore$ Firm B has greater variability, as its coefficient of variation is greater than that of Firm A.

(c) Here, $n_1=586,\ \bar{x}_1=52\text{·}5,\ \sigma_1=10$

$n_2=648,\ \bar{x}_2=47\text{·}5,\ \sigma_2=11$

$$\therefore\quad \bar{x}_{12}=\frac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2}=\frac{586\times 52\text{·}5+648\times 47\text{·}5}{586+648}=\frac{30{,}765+30{,}780}{1234}$$

$$=\frac{61{,}545}{1{,}234}=49\text{·}87=\text{Rs. }49\text{·}9.$$

Again, $d_1=\bar{x}_1-\bar{x}_{12}=52\text{·}5-49\text{·}9=2\text{·}6,$

$d_2=47\text{·}5-49\text{·}9=-2\text{·}4$

$$\therefore\quad \sigma_{12}=\sqrt{\left\{\frac{n_1\sigma_1^2+n_2\sigma_2^2+n_1d_1^2+n_2d_2^2}{n_1+n_2}\right\}}$$

$$=\sqrt{\left\{\frac{586(10)^2+648(11)^2+586(2\text{·}6)^2+648(-2\text{·}4)^2}{586+648}\right\}}$$

$$=\sqrt{\left\{\frac{58600+78408+3962+3733}{1234}\right\}}=\sqrt{\frac{144703}{1234}}=10\text{·}83$$

(Calculation by log table)

Advantages and Disadvantages of Standard Deviation.

Advantages :

(1) Standard deviation is based on all the observations and is rigidly defined.

(2) It is amenable to algebraic treatment and possesses many mathematical properties.

(3) It is less affected by fluctuations of sampling than most other measures of dispersion.

(4) For comparing variability of two or more series, coefficient of variation is considered as most appropriate and this is based on standard deviations and mean.

Disadvantages :

(1) It is not easy to understand and to calculate.

(2) It gives more weight to the extremes and less to the items nearer to mean, since the squares of the deviations of bigger sizes would be proportionately greater than that which are

comparatively small. The deviations 2 and 6 are in the ratio of 1 : 3, but their squares 4 and 36 would be in the ratio of 1 : 9.

Uses of Standard Deviation.

It is the best measure of dispersion, and should be used wherever possible.

More Examples.

(1) From a certain frequency distribution consisting of 18 observations the mean and the standard deviation were found to be 7 and 4 respectively. But on comparing with the original data it was found that a figure 12 was miscopied as 21 in the calculations. Calculate the correct mean and standard deviation. [I. C. W. A. 1965]

We know, $\Sigma x = n\bar{x} = 18 \times 7 = 126$, actual $\Sigma x = 126 - 21 + 12 = 117$.

$$\therefore \quad \text{actual Mean} = \frac{\Sigma x}{n} = \frac{117}{18} = 6{\cdot}5.$$

$$\text{Again} \quad \sigma = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}$$

$$\text{or,} \quad 4 = \sqrt{\frac{\Sigma x^2}{18} - 7^2}$$

$$\text{or,} \quad 16 = \frac{\Sigma x^2}{18} - 49$$

$$\text{or,} \quad \frac{\Sigma x^2}{18} = 16 + 49 = 65$$

$$\text{or,} \quad \Sigma x^2 = 65 \times 18 = 1170.$$

Now actual $\Sigma x^2 = 1170 - (21)^2 + (12)^2 = 1170 - 441 + 144 = 873$.

$$\therefore \quad \text{actual } \sigma = \sqrt{\frac{873}{18} - (6{\cdot}5)^2} = \sqrt{48{\cdot}50 - 42{\cdot}25} = \sqrt{6{\cdot}25} = 2{\cdot}5.$$

(2) The Mean and S. D. of a group of 25 observations were found to be 30 and 3 respectively. After the calculations were made it was found that two of the observations were incorrect, which were recorded as 29 and 31. Find the mean and S. D. if the incorrect observations are excluded. [C. U. B. Com. (Hons.) 1968]

We know, $\Sigma x = n\bar{x} = 25 \times 30 = 750$,

actual $\Sigma x = 750 - (29 + 31) = 690$.

So actual Mean $= \dfrac{\Sigma x}{n} = \dfrac{690}{23} = 30$ (here, $n = 25 - 2 = 23$).

Again, $\sigma^2 = \dfrac{\Sigma x^2}{n} - \left(\dfrac{\Sigma x}{n}\right)^2$

or, $9 = \dfrac{\Sigma x^2}{25} - (30)^2$

or, $\dfrac{\Sigma x^2}{25} = 9 + 900 = 909$

or, $\Sigma x^2 = 25 \times 909 = 22725.$

So actual $\Sigma x^2 = 22725 - (29)^2 - (31)^2$

$= 22725 - 841 - 961 = 20923.$

Now, actual $\sigma = \sqrt{\dfrac{20923}{23} - 900} = \sqrt{909{\cdot}7 - 900} = \sqrt{9{\cdot}7} = 3{\cdot}11.$

(3) The Mean and the Variance calculated from a group of 80 obvervations are 63·2 and 25·93 respectively. If 60 of these observations have mean = 64·8 and S.D. = 4, find the Mean and S. D. of the remaining 20 observations. [I. C. W. A. July 1971]

We assume that the total, *i.e.*, 80 observations have been split up into two groups—

Group A contains 60 ($= n_1$) observations with Mean $\bar{x}_1 = 64{\cdot}8$ and S. D. $\sigma_1 = 4$ and

Group B contains 20 ($= n_2$) observations with mean $\bar{x}_2$ and S. D. σ_2.

Now, for Group A : $n_1 = 60,\ \bar{x}_1 = 64{\cdot}8,\ \sigma_1 = 4$

for Group B : $n_2 = 20,\ \bar{x}_2 = \ ?\ ,\ \sigma_2 = \ ?$

for combined Group : $n_1 + n_2 = 80,\ \bar{x}_{12} = 63{\cdot}2,\ \sigma_{12}{}^2 = 25{\cdot}93.$

Now, from $\bar{x}_{12} = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$ we find,

$$63{\cdot}2 = \frac{60 \times 64{\cdot}8 + 20\bar{x}_2}{80}$$

or, $20\bar{x}_2 = 63{\cdot}2 \times 80 - 60 \times 64{\cdot}8$

or, $20\bar{x}_2 = 5056 - 3888 = 1168.$

$\therefore\ \bar{x}_2 = \frac{1168}{20} = 58{\cdot}4.$

Now, $d_1 = \bar{x}_1 - \bar{x}_{12} = 64{\cdot}8 - 63{\cdot}2 = 1{\cdot}6,$

$d_2 = \bar{x}_2 - \bar{x}_{12} = 58{\cdot}4 - 63{\cdot}2 = -4{\cdot}8.$

Again $\sigma_{12} = \sqrt{\left\{\dfrac{60(4)^2 + 20(\sigma_2)^2 + 60(1{\cdot}6)^2 + 20(-4{\cdot}8)^2}{60 + 20}\right\}}$

$$= \sqrt{\left\{\frac{960 + 20\sigma_2{}^2 + 153{\cdot}6 + 460{\cdot}8}{80}\right\}} = \sqrt{\frac{1574{\cdot}4 + 20\sigma_2{}^2}{80}}$$

$$\text{or,}\quad \sigma_{12}^2=\frac{1574{\cdot}4+20\sigma_2^2}{80}$$

$$\text{or,}\quad 25{\cdot}93=\frac{1574{\cdot}4+20\sigma_2^2}{80}$$

$$\text{or,}\quad 20\sigma_2^2=25{\cdot}93\times80-1574{\cdot}4=2074{\cdot}4-1574{\cdot}4=500$$

$$\text{or,}\quad \sigma_2^2=\frac{500}{20}=25\quad\therefore\quad\sigma_2=5.$$

(4) For two groups of observations, the following results are available :

Group I	Group II
$\Sigma(x-5)=3$	$\Sigma(x-8)=11$
$\Sigma(x-5)^2=43$	$\Sigma(x-8)^2=76$
$n_1=18$	$n_2=17$

Find (correct to 3 significant figures) the mean and the standard deviation of the 35 observations obtained by combinning the two groups. [I. C. W. A. July, 1972]

For Group I : $\text{Mean }(\bar{x}_1)=5+\frac{3}{18}=5+0{\cdot}17=5{\cdot}17.$

$$\text{S. D.}\quad(\sigma_1)=\sqrt{\frac{43}{18}-\left(\frac{3}{18}\right)^2}$$

$$=\sqrt{2{\cdot}389-{\cdot}0289}=\sqrt{2{\cdot}3601}=1{\cdot}537=1{\cdot}54.$$

For group II : $\text{Mean }(\bar{x}_2)=8-\frac{11}{17}=8-{\cdot}65=7{\cdot}35.$

$$\text{S. D.}\quad(\sigma_2)=\sqrt{\frac{76}{17}-\left(\frac{-11}{17}\right)^2}=\sqrt{4{\cdot}4710-{\cdot}4225}$$

$$=\sqrt{4{\cdot}0485}=2{\cdot}013=2{\cdot}01.$$

$$\text{Now,}\quad\bar{x}_{12}=\frac{18(5{\cdot}17)+17(7{\cdot}35)}{18+17}=\frac{218{\cdot}01}{35}=6{\cdot}23.$$

Again $d_1=5{\cdot}17-6{\cdot}23=-1{\cdot}06$;

$d_2=7{\cdot}35-6{\cdot}23=1{\cdot}12$

$$\sigma_{12}=\sqrt{\left\{\frac{18(1{\cdot}54)^2+17(2{\cdot}01)^2+18(-1{\cdot}06)^2+17(1{\cdot}12)^2}{18+17}\right\}}$$

$$=\sqrt{\left\{\frac{42{\cdot}66+68{\cdot}68+20{\cdot}23+21{\cdot}33}{35}\right\}}$$

$$=\sqrt{\frac{152{\cdot}90}{35}}=\sqrt{4{\cdot}37}=2{\cdot}09.$$

Lorenz Curve.

Dispersion can also be studied by the help of Lorenz Curve, after the name of Max. O. Lorenz, the economist statistician. He used this curve to measure the distribution of wealth and income. Lorenz Curve is a percentage cumulative curve in which the percentage of items under review is combined with the percentage of other things as wealth, profits, etc.

For drawing the curve, size of the items and frequencies are both cumulated, and their percentages are calculated for the cumulated values. These percentages are to be plotted in the graph paper. If wealth are equally distributed among the people concerned, the curve would be a straight line joining the extremes of the different scales. This line is known *Line of equal distribution.* If again the distribution is not proportionately equal, it indicates variability, and the curve would be away from the line of equal distribution. The further the curve is away from this line, the greater is the variability. Lorenz Curve does not yield a numerical measure.

Example.

In the table given below is given the number of companies belonging to two area A and B according to the amount of profits earned by them. Draw in the same diagram their Lorenz Curves and interpret them.

Profits earned in Rs. '000	*No. of Companies*	
	Area A	*Area* B
6	6	2
25	11	38
60	13	52
84	14	28
105	15	38
150	17	26
170	10	12
400	14	4

[I. C. W. A. 1964]

Calculation for drawing the Lorenz Curve

Profits			*Area A*			*Area B*		
Profits earned in Rs. '000	*Cumulative profits*	*Cumulative percentage*	*No. of companies*	*Cumulative number*	*Cumulative percentage*	*No. of companies*	*Cumulative number*	*Cumulative percentage*
6	6	0·6	6	6	6	2	2	1
25	31	3·1	11	17	17	38	40	20
60	91	9·1	13	30	30	52	92	46
84	175	17·5	14	44	44	28	120	60
105	280	28·0	15	59	59	38	158	79
150	430	43·0	17	76	76	26	184	92
170	600	60·0	10	86	86	12	196	98
400	1000	100·0	14	100	100	4	200	100

Lorenz Curve

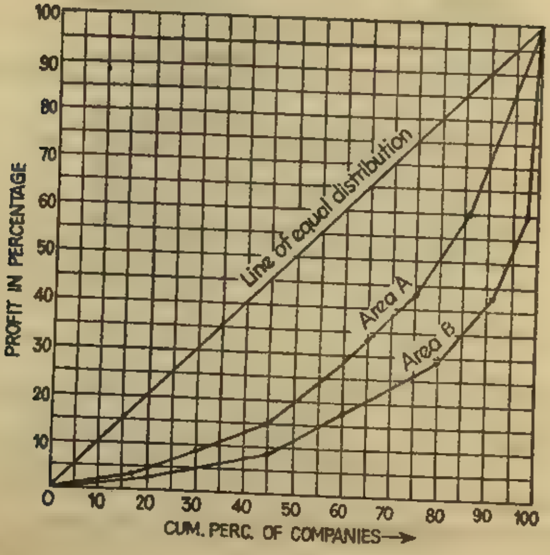


Fig. 35

The curve B is of greater inequality, since it is furthest from the line of equal distribution.

EXERCISE 7

1. Explain the term Dispersion. What purpose does a measure of dispersion serve? Distinguish between absolute and relative measure of dispersion. [Poona, B. Com. 1966]

2. Define Mean Deviation. How does it differ from Standard Deviation? [C. A. Nov. 1968] [I.C.W.A. Jan. 1964]

3. Explain why Standard Deviation is regarded as superior to the other measures of dispersion. What is its chief defect?

4. What are quartiles? How are they used for measuring dispersion?

5. What is coefficient of variation? What purpose does it serve? Also distinguish between 'variance' and 'coefficient of variation'.

6. What is Lorenz Curve? How is it drawn? In what way does it help in studying variation of two or more distributions? Illustrate with the help of an example.

7. If each term is reduced by 10, what effect would this have on (i) the Arithmetic Mean, (ii) the Range, (iii) the Standard Deviation? [C. A. May 1964]

(*Ans.* (i) reduced by 10 ; (ii) & (iii) no change)

8. The weight of 11 forty-year old men were 148, 154, 158, 160, 161, 162, 166, 170, 182, 195 and 236 pounds. If the haviest man is omitted, what is the percentage change in the range?

[C. U. M.Com. 1968] (*Ans.* 46·6)

9. From the following table, compute the Quartile Deviation :

Size :	4—8	8—12	12—16	16—20	20—24	24—28	28—32	32—36	36—40
Freq. :	6	10	18	30	15	12	10	6	2

[C. A. May 1965] (*Ans.* 5·2)

10. The following table gives the monthly wages of 72 workers in a factory. Compute the quartile deviation :

Monthly wages (Rs.) :	12·5—17·5	17·5—22·5	22·5—27·5	27·5—32·5	32·5—37·5	37·5—42·5	42·5—47·5	47·5—52·5	52·5—57·5
No. of workers :	2	22	19	14	3	4	6	1	1

[C. U. M. Com. 1962] (*Ans.* Rs. 5·15)

11. The following table shows the distribution of the maximum loads supported by certain cables produced by a company :

Maximum load (short tons) :	9·3—9·7	9·8—10·2	10·3—10·7	10·8—11·2	11·3—11·7	11·8—12·2	12·3—12·7	12·8—13·2
No. of cables :	2	5	12	17	14	6	3	1

—Find the semi-interquartile range for the above distribution.

(*Ans.* ·50 short tons)

12. Find the mean absolute deviation of the following observations : 2, 4, 9, 16, 20, 10, 14, 13, 8, 10. (*Ans.* 4·12)

13. Find the mean deviation of the following series :

x :	10	11	12	13	14	Total
f :	3	12	18	12	3	48

[C. A. May 1963] (*Ans.* 0·75)

14. Calculate the mean deviation for the following frequency distribution :

No. of colds experienced in 12 months	*No. of persons*	*No. of colds experienced in 12 months*	*No. of persons*
0	15	5	95
1	46	6	82
2	91	7	26
3	162	8	13
4	110	9	2

[I.C.W.A. Jan. 1963] (*Ans.* 1·466)

15. You are given the frequency distribution of 291 workers of a factory according to their average monthly income in 1954-55. Locate median and quartiles. Find also mean deviation about median and hence coefficient of dispersion.

Income group (Rs.)	*No. of workers*	*Income group (Rs.)*	*No. of workers*
Below 50	1	150—170	22
50—70	16	170—190	15
70—90	39	190—210	15
90—110	58	210—230	9
110—130	60	230 and above	10
130—150	46		

(*Ans.* Med. = 120·5 ; Q_1 = 95·78 ; Q_3 = 149·24 ; M.D. = 38·11 ; coeff. disp. = ·73)

16. Calculate the standard deviation from the following table :

Days :	1	2	3	4	5	6	7	8	9	10
Daily Earn. : (Rs.)	1·50	1·00	1·25	2·25	2·00	2·50	3·00	1·50	3·00	2·00

(*Ans.* Rs. 0·66)

17. Find the standard deviation of the following distribution :

x :	4·5	14·5	24·5	34·5	44·5	54·5	64·5
f :	1	5	12	22	17	9	4

[Delhi, B.A. (Hons.) 1969] (*Ans.* 13·25)

18. Calculate the standard deviation from the following data :

Temp. 'C'	*No. of days*	*Temp. 'C'*	*No. of days*
− 40 to 30	10	0 to 10	65
− 30 to 20	28	10 to 20	180
− 20 to 10	30	20 to 30	10
− 10 to 0	42		

[C. A. May 1966] (*Ans.* 14·73°C)

19. The following is the record of goals scored by team A in a football season :

No. of goals scored :	0	1	2	3	4
No. of matches :	1	9	7	5	3

For team B, the average number of goals scored per match was 2·5 with a standard deviation of 1·25 goals.

Find which team may be considered more consistent.

[C. A. Nov. 1963] (*Ans.* team B)

20. The following frequency distributions have been constructed from measurements of heights (in inches) and weights (in lb.) of the same group of adult persons. Which is more variable, height or weight ?

Mid-point (ht.)	*Freq.*
60	10
62	10
64	30
66	30
68	15
70	5
	100

Mid-point (weight)	*Freq.*
84	10
94	10
104	15
114	20
124	15
134	10
144	10
154	10
	100

[I.C.W.A. July 1968] (*Ans.* weight)

21. The number of runs scored by cricketers A and B during a test series consisting of 5 test matches is shown below for each of the 10 innings :

Cricketer A :	5	26	97	76	112	89	6	108	24	16
Cricketer B :	51	47	36	60	58	39	44	42	71	50

Make a comparative study of their batting performance.

(*Ans.* Cricketer B is more consistent scorer)

22. The scores of two batsmen A and B, inter innings during a certain season are as under :

A	32	28	47	63	71	39	10	60	96	14
B	19	31	48	53	67	90	10	62	40	80

—Find which of the batsmen is more consistent in scoring.

[I. C. W. A. Jan. 1970] (*Ans.* Batsman B)

23. From the data given below, state which series is more variable (use standard deviation) :

Variable	*Series A*	*Series B*
10—20	10	18
20—30	18	22
30—40	32	40
40—50	40	32
50—60	22	18
60—70	18	10

[C. A. Nov. 1971] (*Ans.* Series B)

24. The following table gives the distribution of wages in the two branches of an industrial concern :

Monthly wages (Rs.)	*No. of Workers*	
	Branch A	*Branch* B
100—150	167	63
150—200	207	93
200—250	253	157
250—300	205	105
300—350	168	82
	1000	500

Find out the arithmetic mean and the standard deviation for the two branches separately ; state

(i) which branch pays higher average wages per month ;

(ii) which branch has greater variability in wages relative to the average wages ; and

(iii) what is the average monthly wages for the concern as a whole.

[C. A. May 1964] (*Ans.* (i) B ; (ii) A ; (iii) Rs. 227·50)

25. For a set of 100 observations, the sum of deviations from 4 cm. is -11 cm. and the sum of the squares of those deviations is 257 square cm. Find the coefficient of variation.

[I. C. W. A. Jan. 1967] (*Ans.* 41·13)

26. (i) If the first quartile is 118 and semi-interquartile range is 12, find the third quartile. (*Ans.* 142)

(ii) The coefficient of variation is 25 and mean is 20 ; find the standard deviation. (*Ans.* 5)

27. Given the following results relating to two groups containing 20 and 30 observations, calculate the coefficient of variation of all of the 50 observations by combining both the groups :

	Groups	
	I	II
Σx	45	55
Σx^2	118	132

[I. C. W. A. Jan. 1968] (*Ans.* 50)

28. The mean and S. D. calculated from 20 observations are 15 and 10 respectively. If an additional observation 5, left out through oversight, be included in the calculations, find the corrected mean and S. D. [I. C. W. A. Jan. 1969] (*Ans.* 14·52 ; 9·99)

29. The mean income per month of a friendly society of 25 members is Rs. 350 and the standard deviation is Rs. 50. Five more members are admitted to the society and their incomes in Rs. per month are 260, 300, 320, 490 and 590. Find the mean and standard deviation of income for the new group of 30 members.

[I. C. W. A. July 1969] (*Ans.* Rs. 357 ; Rs. 70·65)

30. Given below the frequency distribution of the marks obtained by 90 students. Compute the A.M., Median, Mode and S.D.

Marks :	20—29	30—39	40—49	50—59	60—69	70—79	80—89	90—99
No. of Students :	5	12	15	20	18	10	6	4

[I. C. W. A. June, '76] (*Ans.* 56·5 ; 56 ; 56·64 ; 17·6 marks)

31. (*a*) Define 'Standard Deviation' of ungrouped and grouped data. Discuss its merits and demerits as a measure of dispersion.

(*b*) Calcutate the S.D. of the following observations on a certain variable :

240·12,	240·13,	240·15,	240·12,	240·17,
240·15,	240·17,	240·16,	240·22,	240·21.

[I. C. W. A. June 1976] (*Ans.* ·0302)

32. A company has three establishments E_1, E_2, E_3 in three cities. Analysis of the monthly salaries paid to the employees in the three establishments is given below :

	E_1	E_2	E_3
No. of employees	20	25	40
Average monthly salaries (Rs.)	305	300	340
Standard dev. of monthly salaries (Rs.)	50	40	45

Find the average and the standard deviation of the monthly salaries of all the 85 employees in the company.

[I. C. W. A. Dec. 1976] (*Ans.* Rs. 320 ; Rs. 48·68)

33. Explain with suitable example the term 'dispersion'. Mention some common measures of dispersion and describe the one which you think to be the most important of them.
[I. C. W. A. Dec. 1976]

34. The table below gives the frequency distribution of weights of 80 apples selected at random from a big consignment :

Wt. : (gm.)	110—119	120—129	130—139	140—149	150—159	160—169	170—179	180—189
f :	5	7	12	20	16	10	7	8

(*a*) Draw the cumulative frequency diagram and hence determine the median weight of an apple.

(*b*) Find the coefficient of variation for this distribution.
[I. C. W. A. Dec. 1976] (*Ans.* 147·5 gm. ; 11·75%)

35. An analysis of the monthly wages paid to workers in two firms A and B, belonging to the same industry, gives the following results :

	Firm A	*Firm* B
Number of wage-earners :	550	650
Average monthly wages :	Rs. 50	Rs. 45
S.D. of the distributions of wages :	Rs. $\sqrt{90}$	Rs. $\sqrt{120}$

Answer to the following questions with proper justifications :

(*a*) Which Firm A or B pays out larger amount as monthly wages ?

(*b*) In which Firm, A or B is there greater variability in individual wages ?

(*c*) What are the measures of (i) average monthly wages and (ii) standard deviation in the distribution of individual wages of all workers in the two firms taken together ? [I. C. W. A. June '77]
(*Ans.* (*a*) B ; (*b*) B ; (*c*) Rs. 47·29 ; Rs. 10·64)

36. (*a*) Calculate mean deviation from the median from the following :

Class-intervals :	2—4	4—6	6—8	8—10
Frequencies :	3	4	2	1

(*b*) Calculate the standard deviation of the following distribution :

Age (x) :	20—25	25—30	30—35	35—40	40—45	45—50
No. of persons :	170	110	80	45	40	35

[I. C. W. A. Dec. 1977] (*Ans.* (*a*) 1·4 ; (*b*) 7·94)

37. (*a*) The mean and s.d. of 20 items is found to be 10 and 2 respectively. At the time of checking it was found that one item was incorrect. Calculate the mean and s.d. if

(i) the wrong item is omitted, and

(ii) it is replaced by 12.

(*b*) The means of two samples of sizes 50 and 100 respectively are 54·1 and 50·3 and the standard deviations are 8 and 7. Obtain the s.d. of the sample of size 150 obtained by combining the two samples. [I. C. W. A. Dec. 77]

(*Ans.* (*a*) (i) 10·11, 1·96 ; (ii) 10·2, 1·99.
(*b*) 51·57, 7·56.)

38. (*a*) If the first quartile is 142 and the semi-interquartile range is 18, find the median (assuming the distribution to be symmetrical about mean or median).

(*b*) Prove that the standard deviation is independent of any change of origin but is dependent on the change of scale.

(*c*) Compute the arithmetic mean, standard deviation and the mean deviation about the mean for the following data :

Scores :	4—5	6—7	8—9	10—11	12—13	14—15	Total
f :	4	10	20	15	8	3	60

[I. C. W. A. Dec. 1978] (*Ans.* (*a*) 160, (*c*) 9·233 ; 2·476 ; 2·0311)

39. Calculate the variance of the following distribution (correct up to 3 places after decimal), stating any necessary assumptions :

Height in inches	*No. of men (hundreds)*
under 62	6
62 and under 63	10
63 „ „ 64	10
64 „ „ 65	40
65 „ „ 66	72
66 „ „ 67	78
67 „ „ 68	90
68 „ „ 69	88
69 and over	56
	450

[I.C.W.A. June 1979] (*Ans.* 3·343)

40. The means of two samples of sizes 50 and 100 respectively are 54·4 and 58·5 and the standard deviations are 9 and 11. Obtain the mean and standard deviation of the sample of size 150 obtained by combining the two samples. [I. C. W. A. June 1979] (*Ans.* 57·13 ; 10·56)

41. For a group of 50 boys the mean score and the standard deviation of scores on a test are 59·5 and 8·38 respectively. For a group of 40 girls the same results are 54·0 and 8·23 respectively. Find the mean and the standard deviation of the combined group of 90 children. [C. U. B. Com. (Hons.) 1980] (*Ans.* : 57·06 ; 8·75)

42. The first of the two samples has 100 items with mean 0·23 and S.D. 5·61. If the second sample has 75 items with mean 22·04 and S.D. 1·34 find the mean and variance (square of S.D.) of the sample obtained by combining the two. [I.C.W.A. Dec. 1979] (*Ans.* 9·38 ; 135·20)

43. The means of two samples of sizes 50 and 100 respectively are 54·1 and 50·3 and the S.D. are 8 and 7. Find the S.D. of the sample of size 150 obtained by combining the two samples.

[I.C.W.A. June 1980] (*Ans.* 51·57 ; 7·56)

44. Goals scored by teams A and B in a football match were as follows :

No. of goals scored in a match	*Number of matches*	
	A	B
0	26	18
1	10	8
2	7	5
3	6	6
4	4	3

—Calculate the mean and the S.D. in each case.

[I.C.W.A. Dec. 1979] (*Ans.* For A : 1·1, 1·33 ; For B : 1·2, 1·35)

45. Calculate mean and S.D. of the following data :

Age		*No. of persons dying*
Under	10	15
"	20	30
"	30	53
"	40	75
"	50	100
"	60	110
"	70	115
"	80	125

[I.C.W.A. June 1980] (*Ans.* 35·16 yrs. ; 19·76 yrs.)

9

MOMENTS, SKEWNESS AND KURTOSIS

(A) Moments.

For n observations $x_1, x_2, x_3, \ldots, x_n$, the arithmetic mean of the rth power of deviations taken from an arbitrary constants A, is defined as *rth moment about A* (denoted by m_r).

So, $m_r = \frac{1}{n} \Sigma(x_i - \text{A})^r$

i.e., $m_r = \frac{1}{n} [(x_1 - \text{A})^r + (x_2 - \text{A})^r + (x_3 - \text{A})^r \cdots\cdots + (x_n - \text{A})^r]$

For $r = 1$, $m_1 = \frac{1}{n} \Sigma(x_i - \text{A})$, is known as 1st moment about A

$r = 2$, $m_2 = \frac{1}{n} \Sigma(x_i - \text{A})^2$, ... 2nd ... ... A

$r = 3$, $m_3 = \frac{1}{n} \Sigma(x_i - \text{A})^3$, ... 3rd ... ... A

$r = 4$, $m_4 = \frac{1}{n} \Sigma(x_i - \text{A})^4$, ... 4th ... ... A

and so on. Now it may be noticed (from the 1st moment about A)

$$m_1 = \tfrac{1}{n}\Sigma(x_i - \text{A}) = \tfrac{1}{n}(\Sigma x_i - \Sigma\text{A}) = \tfrac{1}{n}(\Sigma x_i - n\text{A}) = \bar{x} - \text{A}$$

i.e., the 1st moment about $\text{A} = \bar{x} - \text{A}$ (2)

Cases.

Now (i) when A = 0, we find moments about zero. These are called *Raw Moments*. So the rth raw moment is defined as

$$m_r = \frac{\Sigma x_i^r}{n}.$$

For $r = 1$, $m'_1 = \frac{\Sigma x_i}{n} = \bar{x}$ (*arithmetic mean*)

We also find this putting A = 0 in (2)

So, the 1st raw moment is the arithmetic mean.

(ii) When $A=\bar{x}$, we find moments about mean. These are known as *Central Moments.* So rth central moment is defined by

$$m_r' = \frac{1}{n}\Sigma(x_i-\bar{x})^r$$

[dash (′) is used to distinguish from other moment.]

Now for $r=1$, $m_1' = \frac{1}{n}\Sigma(x_i-\bar{x}) = \frac{0}{n} = 0$ as $\Sigma(x_i-\bar{x})=0$

$r=2$, $m_2' = \frac{1}{n}\Sigma(x_i-\bar{x})^2 = \sigma^2$ (square of s.d.)

So we get that 1st central moment is always zero and 2nd central moment is variance σ^2 so that $\sigma = \sqrt{m_2'}$ (s.d. is the square root of 2nd central moment).

Note. The 3rd central moment (m_3') is used to measure skewness and the 4th central moment (m_4') to measure kurtosis. Higher order of moments are of little use.

Example.

For the numbers 2, 4, 6, 8, find the first four moments about 4.

Calculation of Moments

x	$x-4$	$(x-4)^2$	$(x-4)^3$	$(x-4)^4$
2	−2	4	−8	16
4	0	0	0	0
6	2	4	8	16
8	4	16	64	256
Total	4	24	64	288

$$m_1 = \frac{\Sigma(x-4)}{4} = \frac{4}{4} = 1, \qquad m_2 = \frac{\Sigma(x-4)^2}{4} = \frac{24}{4} = 6,$$

$$m_3 = \frac{\Sigma(x-4)^3}{4} = \frac{64}{4} = 16, \qquad m_4 = \frac{\Sigma(x-4)^4}{4} = \frac{288}{4} = 72.$$

Example.

For the numbers 2, 4, 6, 8, find the first four central moments.

Calculation of Moments

x	$(x-5)$	$(x-5)^2$	$(x-5)^3$	$(x-5)^4$
2	-3	9	-27	81
4	-1	1	-1	1
6	1	1	1	1
8	3	9	27	81
20	0	20	0	164

$$\bar{x}(=\text{A.M.})=\frac{\Sigma x}{n}=\frac{20}{4}=5$$

$$m_1'=\frac{\Sigma(x-5)}{n}=\frac{0}{4}=0,\ m_2'=\frac{\Sigma(x-5)^2}{n}=\frac{20}{4}=5,$$

$$m_3'=\frac{\Sigma(x-5)^3}{n}=\frac{0}{4}=0,\ m_4'=\frac{\Sigma(x-5)^4}{n}=\frac{164}{4}=41.$$

Moments from Frequency Distributions.

For the observations $x_1, x_2, \cdots\cdots, x_n$ if $f_1, f_2, \cdots\cdots, f_n$ be the respective frequencies, then the rth moment about A (arbitrary constant) is,

$$M_r=\frac{1}{N}\,\Sigma f_i\,(x_i-A)^r,\ \text{where } \Sigma f_i=N. \qquad \cdots \quad (1)$$

Note. For a grouped frequency distributions, the mid-values of different classes represent the observations $x_1, x_2, x_3, \cdots, x_n$.

Now for $A=0$, we find rth raw moment (or moments about zero),

$$M_r=\frac{1}{N}\,\Sigma f_i\,(x_i)^r\ ;\quad \text{where } \Sigma f=N.$$

For $r=1$, $M_1=\frac{1}{N}\,\Sigma f_i\,x_i=\bar{X}$ (weighted arithmetic mean).

So we again find that the 1st raw moment is A.M.

Again for $A=\bar{X}$, we get rth central moment.

$$M_r'=\frac{1}{N}\,\Sigma f(x_i-\bar{X})^r,\ \text{where } \bar{X}=\frac{\Sigma fx}{N}\ ;\ \Sigma f=N.$$

For $r=1$, $M_1=\frac{1}{N}\Sigma f_i(x_i-A)=\frac{1}{N}\left[\Sigma f_i x_i-\Sigma f_i A\right]$ [from (1)]

$$=\bar{X}-\frac{\Sigma f_i}{N}A=\bar{X}-A \quad (\text{as } \Sigma f_i=N)$$

If $y=\frac{x-c}{d}$, where c and d are constants, then the rth central moment of variate x is equal to d^r times the rth central moment of new variate y. Thus

$$M'_{r(x)}=d^r M'_r(y)$$

As the values of new variate y are small,

$M'_{r(y)}$ may be calculated from the raw moments of y. Now $M'_{r(x)}$ can be calculated multiplying the values of $M'_{r(y)}$ by d^r.

For 1st central moment, *i.e.*, for $A=X$.

We find $M_1'=\bar{X}-\bar{X}=0$

$$r=2, \quad M_2'=\frac{1}{N}\Sigma f_i(x_i-\bar{X})^2=\sigma^2.$$

Thus 2nd central moment is the square of standard deviation and is called the *variance*.

So we find that the 1st central moment is always *zero*, and 2nd central moment is the *variance*.

Example.

Find the first four central moments of the following data :

x :	2	3	5	6
f :	3	2	2	3

Calculation of Moments

x	f	fx	$x-4$	$f(x-4)$	$f(x-4)^2$	$f(x-4)^3$	$f(x-4)^4$
2	3	6	−2	−6	12	−24	48
3	2	6	−1	−2	2	−2	2
5	2	10	1	2	2	2	2
6	3	18	2	6	12	24	48
Total	10	40	0	0	28	0	100

$$\bar{X}=\frac{\Sigma fx}{\Sigma f}=\frac{40}{10}=4.$$

$$M_1'=\frac{\Sigma f(x-4)}{N}=\frac{0}{10}=0,\qquad M_2'=\frac{\Sigma f(x-4)^2}{N}=\frac{28}{10}=2{\cdot}8,$$

$$M_3'=\frac{\Sigma f(x-4)^3}{N}=\frac{0}{10}=0,\qquad M_4'=\frac{\Sigma f(x-4)^4}{N}=\frac{100}{10}=10.$$

Effect of Change of Origin on Moments.

Let the moments about A, *i.e.*, $m_r=\frac{1}{n}\ \Sigma(x_i-A)^r$ is given, and we are to find moments about B, using the moments given.

The moments about B will be $M_r=\frac{1}{n}\ \Sigma(x_i-B)^r$.

Now $x_i-B=(x_i-A)-(B-A)=(x_i-A)-d,$

where $B-A=d,$

or, $(x_i-B)^r=\{(x_i-A)-d\}^r,$

making rth power on both sides

$$=(x_i-A)^r-{}^rc_1(x_i-A)^{r-1}.d+{}^rc_2(x_i-A)^{r-2}d^2-\cdots$$
$$+(-1)^r.d^r \quad \text{(using binomial expansion)}$$

or, $\Sigma(xi-B)^r=\Sigma(xi-A)^r-{}^rc_1d\Sigma(xi-A)^{r-1}$
$$+{}^rc_2d^2\Sigma(xi-A)^{r-2}-\cdots+(-1)^r\Sigma d^r$$

(now taking aggregate (Σ) and multiplying by $\frac{1}{n}$ on both sides)

or, $\frac{1}{n}\Sigma\,(x_i-B)^r=\frac{1}{n}\ \Sigma(x_i-A)^r-{}^rc_1d\ \frac{1}{n}\Sigma(x_i-A)^{r-1}$
$$+{}^rc_2d^2\frac{1}{n}\ \Sigma(x_i-A)^{r-2}-\cdots+(-1)^r\ \frac{1}{n}\cdot nd^r$$

or, $M_r=m_r-{}^rc_1dm_{r-1}+{}^rc_2d^2m_{r-2}-\cdots+(-1)^rd^r$

For $r=1,\quad M_1=m_1-d,$ here $d=B-A$

$r=2,\quad M_2=m_2-2dm_1+d^2$

$r=3,\quad M_3=m_3-3dm_2+3d^2m_1-d^3$

$r=4,\quad M_4=m_4-4dm_3+6d^2m_2-4d^3m_3+d^4.$

Thus we find that the moments about B(M_r) can be expressed as the moments about A(m_r) by the above formulae, where $d=B-A$.

Relation between Central Moments and Non-central Moments.

(*a*) *Expression of Central Moments in terms of Non-central moments.*

Central moment $m_r' = \frac{1}{n}\Sigma(n_i - \bar{x})^r$

Non-central moment about A (say)

$$m_r = \frac{1}{n}\Sigma(n_i - \mathrm{A})^r$$

Now $x_i - \bar{x} = (x_i - \mathrm{A}) - (\bar{x} - \mathrm{A}) = (x_i - \mathrm{A}) - d$,
where $d = \bar{x} - \mathrm{A}$
or, $(x_i - \bar{x})^r = \{(x_i - \mathrm{A}) - d\}^r$

$$= (x_i - \mathrm{A})^r - {}^rc_1(x_i - \mathrm{A})^{r-1}.d + {}^rc_2(x_i - \mathrm{A})^{r-2}d^2 - \ldots\ldots + (-1)^r.d^r$$

or, $\Sigma(x_i - \bar{x})^r = \Sigma(x_i - \mathrm{A})^r - {}^rc_1 d\Sigma(x_i - \mathrm{A})^{r-1}$

$$+ {}^rc_2 d^2\Sigma(x_i - \mathrm{A})^{r-2} - \ldots\ldots + (-1)^2\Sigma d^2$$

or, $\frac{1}{n}\Sigma(x_i - \bar{x})^r = \frac{1}{n}\Sigma(x_i - \mathrm{A})^r - {}^rc_1 d\,\frac{1}{n}\,\Sigma(x_i - \mathrm{A})^{r-1}$

$$+ {}^rc_2 d^2\,\frac{1}{n}\,\Sigma(x_i - \mathrm{A})^{r-2} - \cdots + (-1)^r\frac{1}{n}\Sigma d^r$$

or, $m_r' = m_r - {}^rc_1 d m_{r-1} + {}^rc_2 d^2 m_{r-2} - \ldots\ldots + (-1)^2 d^r$.

Putting $r = 1$, $m_1' = m_1 - d$, here $d = \bar{x} - \mathrm{A}$,

$$m_2' = m_2 - 2m_1 d + d^2,$$
$$m_3' = m_3 - 3m_2 d + 3m_1 d^2 - d^3,$$
$$m_4' = m_4 - 4m_3 d + 6m_2 d^2 - 4m_1 d^3 + d^4.$$

Again we know m_1 (1st moment about A) $= \bar{x} - \mathrm{A}$.

So putting $m_1 = d$,

We find $m_1' = m_1 - m_1 = 0$

$$m_2' = m_2 - 2m_1.m_1 + m_1^2 = m_2 - m_1^2$$
$$m_3' = m_3 - 3m_2 m_1 + 2m_1^3$$
$$m_4' = m_4 - 4m_3 m_1 + 6m_2 m_1^2 - 3m_1^4.$$

(*b*) *Expression of Non-central Moments in terms of Central Moments.*

$(x_i - \mathrm{A})^r = (x_i - \bar{x}) + (\bar{x} - \mathrm{A}) = (x_r - x) + d$, where $d = x - \mathrm{A}$

or, $(x_i - \mathrm{A})^r = (x_i - \bar{x})^r + {}^rc_1(x_i - \bar{x})^{r-1}.d$

$$+ {}^rc_2(x_i - \bar{x})^{r-2}d^2 + \cdots + d^r.$$

or, $\Sigma(x_i - \text{A})^r = \Sigma(x_i - \bar{x})^r + {}^rc_1 d\Sigma(x_i - \bar{x})^{r-1}$
$$+ {}^rc_2 d^2 \Sigma(x_i - \bar{x})^{r-2} + \cdots\cdots + \Sigma d^r$$

or, $\frac{1}{n}\Sigma(x_i - \text{A})^r = \frac{1}{n}\Sigma(x_i - \bar{x})^r + {}^rc_1 d \frac{1}{n}\Sigma(x_i - \bar{x})^{r-1}$
$$+ {}^rc_1 d^2 \frac{1}{n}\Sigma(x_i - \bar{x})^{r-2} + \cdots\cdots + \frac{1}{n}\Sigma d^r$$

or, $m_r = m_r' + {}^rc_1 d m_{r-1}' + {}^rc_2 d^2 m_{r-2}' + \cdots\cdots + d^r$

Taking $r = 1, 2, 3, 4$, we find respectively

$m_1 = m_1' + d$, here $d = x - \text{A}$

$m_2 = m_2' + 2dm_1' + d^2$

$m_3 = m_3' + 3dm_2' + 3d^2 m_1' + d^3$

$m_4 = m_4' + 4dm_3' + 6d^2 m_2' + 4d^3 m_1' + d^4$

Since m_1' (1st central moment) $= 0$, $d = \bar{x} - \text{A} = m_1$

$m_1 = 0 + d = m_1$

$m_2 = m_2' + m_1^2$

$m_3 = m_3' + 3m_2' m_1 + m_1^3$

$m_4 = m_4' + 4m_3' m_1 + 6m_2' m_1^2 + m_1^4$.

Worked out Examples.

(1) The first 3 moments of a distribution about the value 7, calculated from a set of 9 observations are 0·2, 19·4 and -41·0. Find the measures of central tendency and dispersion, and also the third moment about origin. [I. C. W. A. Dec. '75]

Here $m_1 = 0\text{·}2$, $m_2 = 19\text{·}4$, $m_3 = -41\text{·}0$ about A $= 7$. The measures of central tendency and dispersion indicate mean and S.D. (σ)

we know, $m_1 = \bar{x} - \text{A}$

or, $0\text{·}2 = \bar{x} - 7$

or, $\bar{x}$ (mean) $= 7 + 0\text{·}2 = 7\text{·}2$,

or, σ (s.d.) is the square root of 2nd central moment (m_2')

we know, $m'_2 = m_2 - 2dm_1 + d^2$, when $d = \bar{x} - \text{A} = 7\text{·}2 - 7 = 0\text{·}2$

$= 19\text{·}4 - 2(0\text{·}2)(0\text{·}2) + (0\text{·}2)^2 = 19\text{·}4 - 0\text{·}08 + 0\text{·}04$

$= 19\text{·}36$

$\therefore\ \sigma = \sqrt{19\text{·}36} = 4\text{·}4$.

Taking M_3 is the 3rd moment about origin

$\text{M}_3 = m_3 - 3dm_2 + 3d^2 m_1 - d^3$, here $d = 0 - \text{A} = -7$,

$= -41 - 3(-7)\ 19\text{·}4 + 3(-7)^2(0\text{·}2) - (-7)^3$

$= -41 + 407\text{·}4 + 29\text{·}4 + 343 = 738\text{·}8$.

[Alternative Way]

$\frac{1}{n}\Sigma(xi-7)=0\cdot2$ or, $\frac{1}{n}\Sigma xi-7=0\cdot2$ or, $\bar{x}=7+0\cdot2=7\cdot2$

Again $\frac{1}{n}\Sigma(xi-7)^2=19\cdot4$ or, $\frac{1}{n}\Sigma(xi^2-14xi+49)=19\cdot4$

or, $\frac{1}{n}\Sigma xi^2-14\frac{1}{n}\Sigma xi+49=19\cdot4$

or, $\frac{1}{n}\Sigma xi^2-14\bar{x}+49=19\cdot4$

or, $\frac{1}{n}\Sigma xi^2=19\cdot4+14\times7\cdot2-49=19\cdot4+100\cdot8-49=71\cdot2$

$\sigma^2=\frac{1}{n}\Sigma(xi)^2-(\bar{x})^2=71\cdot2-(7\cdot2)^2=19\cdot36,$

$\sigma=\sqrt{19\cdot36}=4\cdot4.$

Next $\frac{1}{n}\Sigma(x_i-7)^3=-41\cdot0$

or, $\frac{1}{n}\Sigma(x_i^3-3x_i^2.7+3.x_i.7^2-343)=-41\cdot0$

or, $\frac{1}{n}\Sigma xi^3-21\cdot\frac{1}{n}\Sigma xi^2+147\frac{1}{n}\Sigma xi-343=-41$

or, $\frac{1}{n}\Sigma xi^3-21\times71\cdot2+147\times7\cdot2-343=-41$

or, $\frac{1}{n}\Sigma xi^3=1495\cdot2-1058\cdot4+343-41=738\cdot8.$

(2) The first four moments about the value 1 are 2·6, 10·2, 43·4 and 192·6 respectively. Find the arithmetic mean and the first four moments about the value 4.

Here $m_1=2\cdot6$, $m_2=10\cdot2$, $m_3=43\cdot4$, $m_4=192\cdot6$ and $A=1$.

We know $m_1=\bar{x}-A$ or, $\bar{x}=m_1+A=2\cdot6+1=3\cdot6.$

Now $d=B-A=4-1=3$, and the required four moments are :

$M_1=m_1-d=2\cdot6-3=-0\cdot4,$

$M_2=m_2-2dm_1+d^2=10\cdot2-2(3)(2\cdot6)+3^2=10\cdot2-15\cdot6+9=3\cdot6,$

$M_3=m_3-3dm_2+3d^2m_1-d^3=43\cdot4-3(3)(10\cdot2)+3.3^2.(2\cdot6)-3^3$

$=43\cdot4-91\cdot8+70\cdot2-27=-5\cdot2,$

$M_4=m_4-4dm_3+6d^2m_2-4d^3m_1+d^4=192\cdot6-4(3)(43\cdot4)$

$+6\cdot3^2(10\cdot2)-4(3)^3(2\cdot6)+3^4$

$=192\cdot6-520\cdot8+550\cdot8-280\cdot8+81=22\cdot8.$

(3) Find the first, second and third central moments of the frequency distribution given below :

Range of expenditure in Rs. per month	*No. of families*
3— 6	28
6— 9	292
9—12	389
12—15	212
15—18	59
18—21	18
21—24	2

[I. C. W. A. Inter. June 1978]

Computation of Moments

Class-interval	*Mid. pt.* x	f	$y=\dfrac{x-13{\cdot}5}{3}$	fy	fy^2	fy^3
3— 6	4·5	28	−3	−84	252	−756
6— 9	7·5	292	−2	−584	1168	−2336
9—12	10·5	389	−1	−389	389	−389
12—15	13·5	212	0	0	0	0
15—18	16·5	59	1	59	59	59
18—21	19·5	18	2	36	72	144
21—24	22·5	2	3	6	18	54
Total	—	1000	—	−956	1958	−3224

Raw moments about y :

$$M_1=\frac{\Sigma fy}{N}=\frac{-956}{1000}=-{\cdot}956$$

$$M_2=\frac{\Sigma fy^2}{N}=\frac{1958}{1000}=1{\cdot}958$$

$$M_3=\frac{\Sigma fy^3}{N}=\frac{-3224}{1000}=-3{\cdot}224.$$

Central moments about y :

$$M'_1=M_1-M_1=0,$$
$$M_2'=M_2-M_1{}^2=1{\cdot}958-(-{\cdot}956)^2=1{\cdot}958-{\cdot}914=1{\cdot}044,$$
$$M_3'=M_3-3M_2M_1+2M_1{}^3=(-3{\cdot}224)-3(1{\cdot}958)(-{\cdot}956)$$
$$+2(-{\cdot}956)^3=-3{\cdot}224+5{\cdot}616-1{\cdot}7480={\cdot}644.$$

Central moments about x :

$$M'_{1(x)} = d \times M'_{1(y)} = 3 \times 0 = 0,$$
$$M'_{2(x)} = d^2 \times M'_{2(y)} = 9 \times 1{\cdot}044 = 9{\cdot}376,$$
$$M'_{3(x)} = d^3 \times M'_{3(y)} = 27 \times {\cdot}644 = 17{\cdot}388.$$

Now mean $(\bar{x}) = c + d\bar{y} = 13{\cdot}5 + 3 \times (-{\cdot}956)$
$$= 13{\cdot}5 - 2{\cdot}868 = 10{\cdot}632.$$

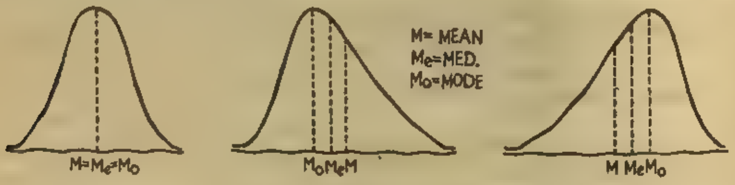
(B) Skewness.

A frequency distribution is said to be 'symmetrical', if the frequencies are distributed symmetrically (or evenly) on either side of an average. When plotted on a graph paper, such distributions will show a normal or ideal curve. In a normal curve, the values of mean, median and mode coincide and the quartiles are equidistant from the median. In such cases, the sum of the deviations measured from mean, median or mode would be zero. A normal curve is a bell-shaped curve, in which the values on either side of an average are symmetrical. In general, frequency distributions are not symmetrical, they are slightly or highly asymmetrical. **Skewness** is opposite to symmetrical. The presence of Skewness indicates that a particular distribution is not symmetrical. The word Skewness literally denotes asymmetry (or lack of symmetry).

Measures of Skewness will not only show the amount of skewness, but also its direction.

A distribution is said to be *positively skewed* when it has a long tail towards the higher values of the variable and *negatively skewed* when the longer tail is present towards the lower values of the variable.

The following figures give us an idea about the shape of symmetrical and asymmetrical curves.



(a) (b) (c)

Fig. 36

Figure 36(*a*) shows the shape of an ideal symmetrical curve. It is bell-shaped and the values of mean, median and mode would be equal.

Figure 36(*b*) indicates a moderately skewed curve. In it the value of the mean would be greater than that of median, which would be also greater than mode. The curve is skewed to the right and is known as *positively skew*.

In Figure 36(*c*), the value of mean would be less than that of median, which would be again *less* than mode. It is skewed to the left and is known as *negatively* skewed.

Test of Skewness.

(1) In a skew distribution, the values of mean, median and mode would not be the same.

(2) Two quartiles would not be equidistant from the median or $(Q_3 - M) - (M - Q_1)$ would not be zero.

(3) When plotted in a graph paper, a skew distribution would not show a bell-shaped curve [as in figure 36(*a*)].

Measures of Skewness.

It has been discussed earlier that the mode is not influenced by the presence of extreme values, the median is influenced by their position only and the mean is influenced by the size of the extreme values. The shape of a frequency distribution as such has an influence on the values of mean, median and mode. For a symmetrical distribution mean, median and mode coincide, but when the distribution is asymmetrical the mean and median move away from the mode towards the extreme values. Mean moves more than median, *i.e.*, for an asymmetrical distribution $M < M_e < M_o$ or $M > M_e > M_o$. Consequently the distance between mean and mode, say, mean − mode may be used to measure skewness.

But such a measure has the following shortcomings :

(i) This measure, being a measure of absolute skewness is always in terms of the unit used in the original observation. So it is not possible to compare the skewness of two distributions which are in different units.

(ii) For identical skewed curves, the same amount of skewness have much different meaning for a distribution of small dispersion than for a distribution of considerably large dispersion.

In order to make valid comparison between the skewness of two or more distributions, some measures are devised which eliminate the above two shortcomings. Such measures, known as coefficient of skewness is a relative measure of skewness obtained by dividing the absolute measures of skewness by some measure of dispersion.

The most widely used measure of skewness as **Pearson's measure of skewness,** which is given by $\dfrac{\text{mean} - \text{mode}}{\text{s.d.}}$.

This skewness will be positive when the skewness is to the right, *i.e.*, when mean > mode, and will be negative when the skewness is to the left, *i.e.*, mean < mode.

For most frequency distribution, it may be difficult to determine the position of mode, while the median may be located satisfactorily. So the empirical relation mean − mode = 3(mean − median) is used to measure the skewness when the distribution is moderately assymmetrical.

So we get $\dfrac{3(\text{mean} - \text{median})}{\text{s.d.}}$ as another measure.

Again skewness may be measured by considering the relative positions of three quartiles. For a symmetrical distribution we get, $M_e - Q_1 = Q_3 - M_e$. For a positively skewed distribution $Q_3 - M_e > M_e - Q_1$, while for negatively skewed distribution $Q_3 - M_e < M_e - Q_1$. Thus $(Q_3 - M_e) - (M_e - Q_1)$ may be taken as an absolute measure of skewness, while again may be put into the relative terms on being divided by $(Q_3 - M_e) + (M_e - Q_1) = Q_3 - Q_1$. So the relative measure is $\dfrac{(Q_3 - M_e) - (M_e - Q_1)}{Q_3 - Q_1} = \dfrac{Q_3 + Q_1 - 2M_e}{Q_3 - Q_1}$. This is known as Bowley's Measure.

All the above measure of skewness has two desirable properties which any measure of skewness should have. They are equal to zero, when the distribution is symmetrical and are all pure numbers.

It has been found that all the measures of skewness vary between −1 and +1.

So we find

(*a*) ***First measure of skewness*** :

Karl Pearson's coefficient of skewness $= \dfrac{\text{mean} - \text{mode}}{\text{s.d.}}$

or, (when mode is ill-defined) $= \dfrac{3(\text{mean} - \text{median})}{\text{s.d.}}$

(*b*) ***Second measure of skewness*** :

Bowley's measure of coefficient of skewness $= \dfrac{Q_3 + Q_1 - 2\,\text{median}}{Q_3 - Q_1}$.

(*c*) ***Moment measure*** :

Coefficient of skewness $= \dfrac{(m_3')^2}{(m_2')^3}$, where m_2', m_3' are the second and third central moments respectively. Coefficient of skewness is represented by β_1 (*read* as 'Beta one')

$$\therefore \quad \beta_1 = \frac{(m_3')^2}{(m_2')^3}.$$

Some statisticians use $\beta_1 = \dfrac{m_3'}{(m_2')^{\frac{3}{2}}}$.

It has been found that the value of Bowley's measure lies between -1 and $+1$.

Note : (i) β_1 will be zero only in case of symmetrical distribution. The greater value of β_1 indicates that the distribution will be more curved.

(ii) The positive value of β_1 means the distribution is positively skewed. Again the negative value of β_1 indicates the distribution is negatively skewed.

(iii) If mean > median > mode, then the distribution is positively skewed. Again if mean < median < mode, then the distribution is negatively skewed.

Example.

Comment on the following results of averages of any distribution :

(i) A.M. is 10, median is 11.

(ii) A.M. is 15, median is 12.

(iii) Mode is 11, median is 13.

(iv) Median is 10, A.M. is 14.

(v) Median is 12, Mode is 13.

(i) A.M. (10) < median (12),
the distribution is negatively skewed.

(ii) A.M. (15) > median (12),
the distribution is positively skewed.

(iii) Median (13) > mode (11),
the distribution is positively skewed.

(iv) A.M (14) > median (10),
the distribution is positively skewed.

(v) Median (12) < mode (13),
the distribution is negatively skewed.

Example.

Compute Karl Pearson's coefficient of skewness from the following data :

variable	*frequency*	*variable*	*frequency*
20·5—23·5	17	29·5—32·5	194
23·5—26·5	193	32·5—35·5	27
26·5—29·5	399	35·5—38·5	10

[Delhi, B. Com. 1953]

Calculation of coefficient of skewness

Variable	*Mid. pt.* x	f	d	$d' = \frac{d}{3}$	fd'	fd'^2
20·5—23·5	22	17	−6	−2	−34	68
23·5—26·5	25	193	−3	−1	−193	193
26·5—29·5	28	399	0	0	0	0
29·5—32·5	31	194	3	1	194	194
32·5—35·5	34	27	6	2	54	108
35·5—38·5	37	10	9	3	30	90
Total	—	840	—	—	51	653

Formula for required coefficient of skewness :

$$\frac{\text{mean} - \text{mode}}{\text{s.d.}(\sigma)}$$

Now, $\text{mean} = A + \frac{\Sigma fd'}{\Sigma f} \times i = 28 + \frac{51}{840} \times 3 = 28 + {\cdot}182 = 28{\cdot}182.$

Mode lies in the class (26·5—29·5),

$l = 26{\cdot}5,\ f_0 = 193,\ f_1 = 399,\ f_2 = 194,\ i = 3$

$$\therefore\quad \text{Mode} = 26{\cdot}5 + \frac{399 - 193}{2 \times 399 - 193 - 194} \times 3$$

$$= 26{\cdot}5 + \frac{206}{411} \times 3$$

$$= 26{\cdot}5 + 1{\cdot}5 = 28.$$

$$\text{s.d. } (\sigma)=\sqrt{\left\{\frac{\Sigma fd'^2}{\Sigma f}-\left(\frac{\Sigma fd'}{\Sigma f}\right)^2\right\}}\times i$$

$$=\sqrt{\left\{\frac{653}{840}-\left(\frac{51}{480}\right)^2\right\}}\times 3$$

$$='8784\times 3=2'6352=2'64 \text{ (calculation by log table)}$$

$$\text{Now, coefficient of skewness}=\frac{\text{mean}-\text{mode}}{\text{s.d.}}$$

$$=\frac{28'18-28}{2'64}$$

$$=\frac{'18}{2'64}='068.$$

Example.

For a moderately skewed data A.M. = 100, coefficient of variation = 35, Karl Pearson's coefficient of skewness = 0'2, find mode and median.

$$\text{coefficient of variation}=\frac{\text{S.D.}}{\text{A.M.}}\times 100$$

$$\text{or,}\quad 35=\frac{\text{S.D.}}{100}\times 100$$

$$\therefore \text{ S.D.}=35.$$

$$\text{Karl Pearson's coefficient of skewness}=\frac{\text{A.M.}-\text{Mode}}{\text{S.D.}}$$

$$\text{or,}\quad 0'2=\frac{100-\text{mode}}{35}$$

$$\text{or,}\quad 100-\text{mode}=35\times 0'2$$

$$\text{or,}\quad 100-\text{mode}=7'0$$

$$\text{or,}\quad \text{mode}=100-7=93.$$

$$\text{We know, A.M.}-\text{mode}=3(\text{A.M.}-\text{median})$$

$$\text{or,}\quad 100-93=3(100-\text{median})$$

$$\text{or,}\quad 7=300-3\text{ median}$$

$$\text{or,}\quad 3\text{ median}=293$$

$$\text{or,}\quad \text{median}=\frac{293}{3}=97'67.$$

Example.

The second and third central moments of four numbers are 5 and 0 respectively. Find the coefficient of skewnes (β_1) by moment measure.

$$\beta_1=\frac{(m_3')^2}{(m_2')^3}=\frac{0}{25}=0$$

Here β_1 is zero, it means that the distribution is symmetrical.

(C) Kurtosis.

The expression **Kurtosis** indicates,. whether a particular distribution is more flat-topped or more peaked than the normal distribution. The normal curve (or bell-shaped curve) is called *Mesokurtic.* The curve which is more flat-topped than the normal curve is known as *Platykurtic* and the curve which is more peaked than the normal curve is known as *Leptokurtic.* From the given Figure 37, the idea of the curves will be clear.

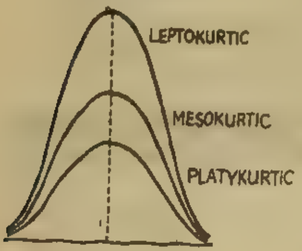


Fig. 37

Measures of Kurtosis.

The coefficient β_2 (read as 'Beta two') is used for the measure, where $\beta_2=\frac{m_4'}{(m_2')^2}$. The standard value of β_2 is 3. In a normal (or mesokurtic) curve, β_2 is 3. If $\beta_2 < 3$, the curve is more flat-topped, *i.e.*, the curve is platykurtic. If again $\beta_2 > 3$, the curve is leptokurtic, *i.e.*, the curve is more peaked.

It may be mentioned here that the knowledge of central moments is utilised in finding out kurtosis. Kurtosis is mainly used in biological studies.

Dispersion, Skewness and Kurtosis.

Dispersion indicates the scatteredness of items round a central value. In skewness we find the extent of deviations below or above an average. Measures of skewness give the shape of the series and size of variation on either side of an average. Kurtosis studies the concentration of items at the central part of a series.

Example.

The first four central moments of a distribution are 0, 2·5, 0·7 and 18·75. Test the skewness and kurtosis of the distribution.

Coefficient of skewness $(\beta_1) = \frac{(m_3')^2}{(m_2')^3} = \frac{(\cdot 7)^2}{(2\cdot 5)^3} = +0\cdot 031.$

Since β_1 is $+ve$, the distribution is positively skewed. For Kurtosis, we are to calculate $\beta_2 = \frac{m_4'}{(m_2')^2}$.

Now $\beta_2 = \frac{18\cdot 75}{(2\cdot 5)^2} = 3.$

Since $\beta_2 = 3$, the distribution is normal, *i.e.*, the curve is mesokurtic.

Example.

The first four central moments of a distribution are 0, 2·5, 0·7 and 18·75. Examine the skewness and kurtosis of the distribution.

Here $\beta_1 = \frac{(m_3')^2}{(m_2')^3} = \frac{(0\cdot 7)^2}{(2\cdot 5)^3} = +0\cdot 031.$

As β_1 is positive, so the distribution is positively skewed.

Again $\beta_2 = \frac{m_4'}{(m_2')^2} = \frac{18\cdot 75}{(2\cdot 5)^2} = 3.$

Since β_2 is 3, *i.e.*, the distribution is normal, so the curve is mesokurtic.

Example.

First four central moments are 0, 6, 12, 120. Examine the skewness and kurtosis.

Here $m_1' = 0,\ m_2' = 6,\ m_3' = 12,\ m_4' = 120$

$$\beta_1(=\text{skewness}) = \frac{(m_3')^2}{(m_2')^3} = \frac{12^2}{6^3} = \frac{144}{216} = +0\cdot 667$$

$\therefore$ The distribution is positively skewed.

Agaio $\beta_2(=\text{kurtosis}) = \frac{m_4'}{(m_2')^2} = \frac{120}{6^2} = \frac{120}{36} = 3\cdot 33.$

Here $\beta_2 > 3$, *i.e.*, the distribution is leptokurtic.

EXERCISE 8

1. Define moments. Establish the relationship between the moments about mean and terms of moments about any arbitrary point and vice versa. [I.C.W.A. Inter, June '77 (N. S.)]

2. Find the first three central moments of 5, 8, 12, 14, 16.

(*Ans.* 0, 16, −22)

3. Find the first three raw moments of 5, 8, 12, 14, 16.

(*Ans.* 11, 137, 1841)

4. The first two moments of a distribution about the value 5 of the variable are 2 and 20. Find the mean and variance.

[I.C.W.A. Inter, June '77] (*Ans.* 7 ; 16)

5. Find the first four central moments of the following data :

x :	2	3	4	5	6
f :	1	3	7	3	1

(*Ans.* 0 ; 0·933 ; 0 ; 2·533)

6. Find the first three moments above mean of the following data :

x :	2	3	4	5
f :	3	2	2	2

(*Ans.* 0 ; 1·45 ; 0)

7. The first four moments about 1 are 2·6, 10·2, 43·4 and 192·6 respectively. Find A.M. and also find the first four moments about 4.

(*Ans.* 3·6 ; −·4, 3·6, −5·2, 22·8)

8. The first three moments about 3 are respectively 2, 10, 30. Find the first three raw moments. Show also that the variance of the distribution is 6. [I.C.W.A. Jan. 1964] (*Ans.* 5, 31, 201)

9. The first four moments about 2 are 1, 2·5, 5·5 and 16. Find the first four moments about A.M. and zero.

(*Ans.* 0, 1·5, 0, 6 ; 3, 10·5, 40·5, 168)

10. Comment on the following average values of a distribution :

(*i*) Median is 21, mode is 12

(*ii*) A.M. is 12, median is 14

(*iii*) A.M. is 10, mode is 8

(*iv*) Mode is 15, median is 12

(*v*) Mode is 12, A.M. is 10

(*vi*) A.M. is 10, median is 10, mode is 10

[*Ans.* (*i*) + (*ii*) − (*iii*) + (*iv*) − (*v*) −vely skewed, (*vi*) symmetrical distribution]

11. Find the Karl Pearson's coefficient of skewness from the following table :-

x :	10	11	12	13	14	15
f :	2	4	10	8	5	1

(*Ans.* ·357)

12. Find the Karl Pearson's coefficient of skewness of the following distribution table :

Age	10—12	12—14	14—16	16—18	18—20	20—22	22—24
No. of students	4	10	16	30	20	14	6

(*Ans.* ·07)

13. Compute the Bowley's coefficient of skewness from the following data :

Marks	No. of students
0—10	25
10—20	15
20—30	20
30—40	15
40—50	20
50—60	30
60—70	65
70—80	50

(*Ans.* −0·473)

14. Find β_1 and β_2 of the data given in question 5 and comment on it. (*Ans.* 0, symmetry ; 2·908, platykurtic)

15. Find β_1 and β_2 of the data given in question 9 and comment on it. (*Ans.* 0, symmetry ; 2·67, platykurtic)

16. Compute quartile deviation and coefficient of skewness given in the following values :

Median = 18·8 cm., Q_1 = 14·6 cm., Q_3 = 25·2 cm.

(*Ans.* 5·3 cm. ; 0·2)

17. Calculate first four moments from the following data :

x :	0	1	2	3	4	5	6	7	8
f :	5	10	15	20	25	20	15	10	5

Also calculate the values of β_1 and β_2 and hence comment on the nature of distribution.

[*Ans.* 0 ; 4 ; 0 ; 37·6 ; $\beta_1 = 0$ (Symmetrical) ; $\beta_2 < 3$ (Platykurtic)]

18. Find the second, third and fourth central moments of the frequency distribution given below. Hence find (i) a measure of skewness (β_1) and (ii) a measure of kurtosis (β_2).

Class-limits :	110·0—114·9	115·0—119·9	120·0—124·9	125·0—129·9	130·0—134·9	135·0—139·9	140·0—144·9
Frequency :	5	15	20	35	10	10	5

[I.C.W.A. Inter, June '76]

(*Ans.* 54 ; 100·5 ; 7827 ; $\beta_1 = ·2533$; $\beta_2 = -·3158$)

19. Calculate Karl Pearson's coefficient of skewness from the following data :

Monthly salary (*Rs.*)		*No. of workers*
below	80	12
„	90	30
„	100	65
„	110	107
„	120	157
„	130	202
„	140	222
„	150	230

(*Ans.* 0·248)

20. Using moments, calculate a measure of relative skewness and a measure of relative kurtosis for the following distribution and comment :

Monthly wages (*Rs.*)				*No. of workers*
70	but	below	90	8
90	„	„	110	11
110	„	„	130	18
130	„	„	150	9
150	„	„	170	4

[*Ans.* 0·08 ; 2·306 (platykurtic)]

10

CORRELATION AND REGRESSION

(A) Correlation.

Meaning : In the previous chapters, we have discussed problems relating one variable only. We have observed how measures of central tendency, measures of dispersion and skewness are calculated for comparison and analysis. We have also seen how series are represented by diagrams and charts. In practice, we face a large number of problems involving the use of *two or more* variables.

If two sets of variables vary in such a way that changes of one set are related by changes in the other, then these sets are said to be *correlated*. For *example*, there is a relation between income and expenditure of a common family, heights and weights of a group of persons, rainfall and production of few commodities, age of husband and age of wife, marks obtained by a group of students in two different subjects, price and demand of a commodity, etc. It is likely that if the income of a common family increases, expenditure also increases of that family. Again in general, with the increase of height of a person, the weight also increases. It may be mentioned here that the two sets of variables should be correlated or interdependent to each other.

If the number of good cricket players in India increases and the production of jute in Bangladesh increases, we cannot say that the phenomena under consideration are related to each other, or there is any correlation in between. In other words, the variables are *uncorrelated*.

Definition : Correlation means the relationship between two variables where with the changes in the values of one variable, the values of other variable also change.

Correlation is also known as *Co-variation*.

(a) *Positive, Negative and Zero Correlation.*

Positive Correlation : A correlation is said to be positive, when high values of one variable are accompanied by the high values of the other, and, that low values of one are accompanied by low values of the other. In positive correlation we find that the two sets of variables always vary in the same direction.

Negative Correlation : In this case high values of one variable are accompanied by the low values of the other. On the other hand, if the values of two variables change in opposite directions, then it is negative correlation.

Zero Correlation : When some high values are accompanied by low values and others are accompanied by high values. In this case, the paired observations are randomly scattered. The variables are also known to be uncorrelated.

***Example*:**

Positive Correlation		*Negative Correlation*	
x	y	x	y
4	10	10	20
6	14	15	18
8	18	20	16
10	22	25	14
12	26	30	12

(*b*) *Simple, Partial and Multiple Correlation.*

The distinction is based upon the number of variables used.

When only two variables are studied it is a *simple* correlation. When there are three or more variables for comparison it is *multiple* or *partial* correlation. In multiple correlation, three or more variables are studied simultaneously. For example, the yield of a commodity is correlated with the amount of rainfall and the amount of fertilizers used.

(*c*) *Linear and Non-linear (Curvilinear) Correlation.*

Correlation will be *linear*, if the variations in the values of two variables are in a constant ratio.

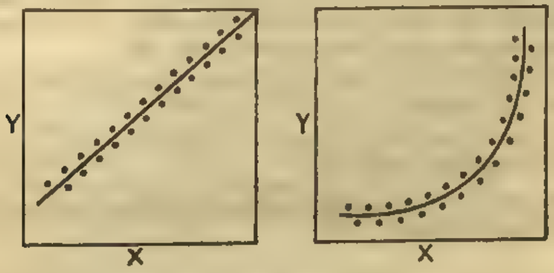
Example :

x	y
10	50
15	75
20	100
25	125
30	150

It is clear that the ratio of change between the two variables is the same. If such variables are plotted on a graph paper, we will find a straight line.

Conversely, if the variations of values of the variables do not bear a constant ratio, we will find *non-linear* or *curvilinear* correlation.

Difference between linear and non-linear correlation will be clear from the following diagrams :



Positive Linear Correlation *Non-Linear Correlation*

Fig. 38

Methods of Studying Correlation.

The following are some of the important methods of studying correlation :

(1) *Scatter Diagram Method,*

(2) *Karl Pearson's Coefficient of Correlation,*

(3) *Rank Method.*

(1) ***Scatter Diagram Method :***

Scatter diagram is a special type of dot chart. For this method the given data are plotted in a graph paper in form of dots. For each pair of x and y values, we put a dot (or point) and thus we obtain many dots equal in number of observations. If now these

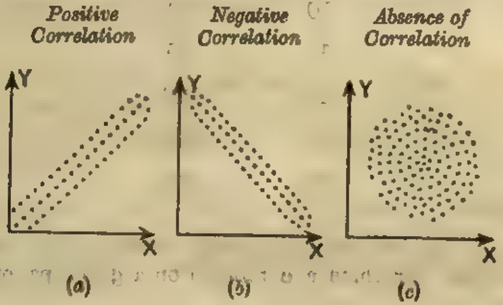


Fig. 39

plotted dots (or points) show some trend either upward or downward, then the two variables (x and y) are said to be correlated, or otherwise not correlated. If again the trend of the points is upward moving from lower left-hand corner to upper right-hand corner, then correlation is positive [$\gamma = +1$).(γ is coefficient of correlation]. On the other hand, if movement is reverse, *i.e.*, dots more from upper left-hand corner to lower right-hand corner, then correlation is *negative* ($\gamma = -1$). The idea will be clear from the diagrams (Fig. 39).

In Fig. 39(a), the values of the two variables move in the same direction, the correlation is positive and $\gamma = 1$.

In Fig. 39(b), we find negative correlation and $\gamma = -1$, as the values move in reverse direction.

In Fig. 39(c), we do not get any trend line and hence it shows the absence of correlation and $\gamma = 0$.

Example.

Given the following pairs of value of the variable X and Y :

X :	3	5	6	10	14	16
Y :	6	5	8	12	17	20

(a) Make a Scatter diagram.

(b) Do you think that there is any correlation between the variables X and Y ? Is it positive or negative ?

Hints : After drawing the scatter diagram it will be seen that the correlation exists and it is positive, *i.e.*, $\gamma = +1$.

Note. In the scatter diagram if the plotted points (or dots) are close to each other, then there would be a high degree of correlation. In such case, a straight line may also be made to pass through such points.

Correlation Coefficient.

In the Scatter Diagram of fig. 39(*c*), let the origin be shifted to O′ whose co-ordinates are $(\bar{x}, \bar{y})$ with respect to the original axes OX and OY of the rectangular co-ordinate and let the two new axes be O′X′ and O′Y′. $\bar{x}$ and y are means of x and y variates respectively.

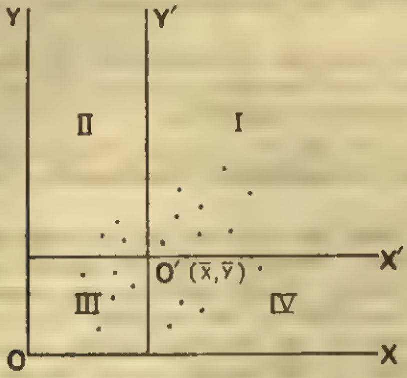


Fig. 40

Now the co-ordinates (x, y) of any point P (say) with reference to the original axes, will be x' and y' with reference to the new axes O′X′ and O′Y′ respectively when $x_i' = x_i - \bar{x}$ and $y_i' = y_i - \bar{y}$.

The points x_i', y_i' in the Scatter Diagram are now distributed over the four quadrants I, II, III and IV (see Fig. 40).

Now, since x_i' and y_i' are deviations from $\bar{x}$ and $\bar{y}$ respectively, then,

(*i*) In the quadrant I, the values of both x' and y' are positive and so also their product $x'y'$ and hence $\Sigma x'y'$ is positive.

(*ii*) In the quadrant II, the values of x' are negative and those of y' are positive, so their product $x'y'$ is negative and hence $\Sigma x'y'$ is negative.

(*iii*) Now in the quadrant III, the values of both x' and y' are negative and hence their product $x'y'$ is positive. $\Sigma x'y'$ is also positive.

(*iv*) Lastly in quadrant IV, the values of x' are positive and that of y' are negative, so that $x'y'$ is negative and hence $\Sigma x'y'$ is negative.

When the correlation is positive, the general tendency of the points is to lie in the 1st and 3rd quadrants, so that the sum of $x'y'$

of all points in 1st and 3rd quadrants is greater than the sum of $x'y'$ of all the points in 2nd and 4th quadrants and hence the sum of all $x'y'$ becomes a positive quantity.

Similarly in case of negative correlation the concentration of the points in 2nd and 4th quadrants is greater than that of the points in the other quadrants so that the sum of all $x'y'$ becomes a negative quantity.

Lastly, when there is no correlation, *i.e.*, when the points are evenly distributed in all the four quadrants, so that the sum of all $x'y'$ will be nearly zero or equal to zero.

Thus the sum of all $x'y'$, *i.e.*, $\Sigma x'y' = \Sigma(x-\bar{x})(y-\bar{y})$ seems to be a natural measure of correlation. But there are two shortcomings in this measure.

In the first case, the value of $\Sigma x'y'$ depends on the number of pairs of observations. Secondly being an absoluto measure it would be in terms of unit in which variables are measured and also it depends on the variability of the variables. So the measure $\Sigma x'y' = \Sigma(x-\bar{x})(y-\bar{y})$ may be put into the relative terms and the above difficulties overcome by dividing by n and the product of standard diviations σ_x and σ_y. The ratio is Pearson's product moment correlation coefficient γ, where

$$\gamma = \frac{\frac{1}{n}\Sigma x'y'}{\sigma_x\sigma_y} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n.\sigma_x\sigma_y} \quad \ldots \qquad \ldots \text{(i)}$$

where n = number of pairs of observations,

$$\sigma_x = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}},$$

$$\sigma_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n}}.$$

The numerator $\frac{1}{n}\Sigma x'y' = \frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y})$ is called the *Covariance* between the two variables x and y and is written as $\text{cov}(x, y)$. This $\text{cov}(x, y)$ has analogy with the term variance of a single variable. For by definition,

$$\text{var}(x) = \frac{1}{n}\Sigma(x-\bar{x})^2 = \frac{1}{n}\Sigma(x-\bar{x})(x-\bar{x})$$

$$\text{var}(y) = \frac{1}{n}\Sigma(y-\bar{y})^2 = \frac{1}{n}\Sigma(y-\bar{y})(y-\bar{y}).$$

Now, $\text{cov}(x, y) = \frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y}).$

So we may also write,

$$\gamma = \frac{\text{cov}\,(x, y)}{\sqrt{\text{var}\,(x)}\,\sqrt{\text{var}\,(y)}} \quad \cdots \quad \text{(ii)}$$

since $\sigma_x = \sqrt{\text{var}\,(x)}$ and $\sigma_y = \sqrt{\text{var}\,(y)}$.

Again,

$$\begin{aligned}
\text{cov}\,(x, y) &= \frac{1}{n}\,\Sigma(x-\bar{x})(y-\bar{y})\\
&= \frac{1}{n}\,\Sigma(xy - x\bar{y} - \bar{x}y + \bar{x}\,\bar{y})\\
&= \frac{1}{n}\,[\Sigma xy - \bar{y}\Sigma x - \bar{x}\Sigma y + \Sigma\bar{x}\,\bar{y}]\\
&= \frac{1}{n}\,\Sigma xy - \bar{y}\cdot\frac{1}{n}\,\Sigma x - \bar{x}\cdot\frac{1}{n}\,\Sigma y + \frac{1}{n}\,\Sigma\bar{x}\,\bar{y}\\
&= \frac{1}{n}\,\Sigma xy - \bar{y}.\bar{x} - \bar{x}\,\bar{y} + \bar{x}\,\bar{y} = \frac{1}{n}\,\Sigma xy - \bar{x}\,\bar{y}
\end{aligned}$$

$$\begin{aligned}
\text{var}\,(x) &= \frac{1}{n}\,\Sigma(x-\bar{x})^2 = \frac{1}{n}\,\Sigma(x^2 - 2x\bar{x} + \bar{x}^2) = \frac{1}{n}\,[\Sigma x^2 - \Sigma 2x\bar{x} + \Sigma\bar{x}^2]\\
&= \frac{1}{n}\,\Sigma x^2 - \frac{1}{n}\,\Sigma 2x\bar{x} + \frac{1}{n}\,\Sigma\bar{x}^2 = \frac{1}{n}\,\Sigma x^2 - 2\bar{x}\cdot\frac{1}{n}\,\Sigma x + \frac{1}{n}\cdot x\cdot\bar{x}^2\\
&= \frac{1}{n}\,\Sigma x^2 - 2\bar{x}\cdot\bar{x} + \bar{x}^2 = \frac{1}{n}\,\Sigma x^2 - \bar{x}^2.
\end{aligned}$$

Similarly, $\text{var}\,(y) = \frac{1}{n}\Sigma y^2 - \bar{y}^2$.

The correlation coefficient may also be written as

$$\gamma = \frac{\frac{1}{n}\,\Sigma xy - \bar{x}\,\bar{y}}{\sqrt{\frac{1}{n}\,\Sigma x^2 + \bar{x}^2}\,\sqrt{\frac{1}{n}\,\Sigma y^2 - \bar{y}^2}} \quad \cdots \quad \text{(iii)}$$

$$= \frac{\frac{1}{n}\,\Sigma xy - \frac{\Sigma x}{n}.\frac{\Sigma y}{n}}{\sqrt{\frac{1}{n}\,\Sigma x^2 - \left(\frac{\Sigma x}{n}\right)^2}\,\sqrt{\frac{1}{n}\,\Sigma y^2 - \left(\frac{\Sigma y}{n}\right)^2}} \quad \cdots \quad \text{(iv)}$$

since $\bar{x} = \frac{\Sigma x}{n}$, $\bar{y} = \frac{\Sigma y}{n}$

$$= \frac{n\Sigma xy - \Sigma x\,\Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\,\sqrt{n\Sigma y^2 - (\Sigma y)^2}} \quad \cdots \quad \text{(v)}$$

For computational purpose, the last form is generally used.

(2) *Karl Pearson's Coefficient of Correlation :*

By this coefficient (popularly known as Personian Coefficient of Correlation) we can measure the extent of relationship between two sets of data. If the correlation is *perfect*, then the coefficient is *unity* or 1. Of course, correlation may be positive or negative according to its nature. If again the coefficient is *zero*, then there is no correlation.

(*a*) For deviations taken from actual A.M.

Pearsonian coefficient of correlation is found by the formula :—

$$\gamma = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}, \text{ where}$$

$x = X - \bar{X}$, *i.e.*, deviation from A.M. of X-series,

$y = Y - Y$, *i.e.*, deviation from A.M. of Y-series.

Example.

Find the coefficient of correlation between X and Y.

X :	1	2	3	4	5	6	7	8	9	10	11
Y :	4	7	10	13	16	19	22	25	28	31	34

		x	y	x^2	y^2	xy
X	Y	$(= X - \bar{X})$	$(= Y - \bar{Y})$			
1	4	−5	−15	25	225	75
2	7	−4	−12	16	144	48
3	10	−3	−9	9	81	27
4	13	−2	−6	4	36	12
5	16	−1	−3	1	9	3
6	19	0	0	0	0	0
7	22	1	3	1	9	3
8	25	2	6	4	36	12
9	28	3	9	9	81	27
10	31	4	12	16	144	48
11	34	5	15	25	225	75
Total 66	209	–	–	110	990	330
				$(= \Sigma x^2)$	$(= \Sigma y^2)$	$(= \Sigma xy)$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{66}{11} = 6 \; ; \; \bar{Y} = \frac{\Sigma Y}{x} = \frac{209}{11} = 19.$$

Now $\gamma=\dfrac{\Sigma xy}{\sqrt{\Sigma x^2\times\Sigma y^2}}=\dfrac{330}{\sqrt{110\times990}}=\dfrac{330}{\sqrt{108900}}=\dfrac{330}{330}=1.$

The correlation of the variables X and Y is perfectly positive.

Example.

From the following results, find the value of coefficient of correlation :

$$\sum_{i=1}^{9}(X_i-\bar{X})^2=60\ ;\quad \sum_{i=1}^{9}(Y_i-\bar{Y})^2=60\ ;\quad \sum_{i=1}^{9}(X_i-\bar{X})(Y_i-\bar{Y})=57.$$

We know, $\Sigma x^2=\Sigma(X_i-\bar{X})^2=60,\ \Sigma y^2=\Sigma(Y-\bar{Y})^2=60,$

$$\Sigma xy=\Sigma(X-\bar{X})(Y-\bar{Y})=57,\ \text{here } n=9.$$

Now from $\gamma=\dfrac{\Sigma xy}{\sqrt{\Sigma x^2\times\Sigma y^2}},$

we get $\gamma=\dfrac{57}{\sqrt{60\times60}}=\dfrac{57}{60}=0{\cdot}95.$

(*b*) For deviations taken from Assumed Mean

If the actual means of the variables are in fractions, then the calculation by the above method will be too lengthy. So we shall use assumed mean for taking deviations. In this case, the following formula will be used.

$$\gamma_{xy}=\gamma d_xd_y=\frac{n\Sigma d_xd_y-\Sigma d_x\Sigma d_y}{\sqrt{n\Sigma d_x^2-(\Sigma d_x)^2}\ .\ \sqrt{n\Sigma d_y^2-(\Sigma d_y)^2}}\quad\cdots\quad(\text{vi})$$

where $d_x=\dfrac{x-A}{C}$ and $d_y=\dfrac{y-B}{D}$, where A, B, C and D are arbitrary constants (C and D are positive) and n is the number of pairs of observations.

Note. For grouped frequency distribution, same formula is to be used considering frequency only.

Example.

Find the coefficient of correlation of the following data :

X :	1	2	3	4	5	6	7	8	9	10
Y :	46	42	38	34	30	26	22	18	14	10

X	Y	d_x $= X-5$	d_y $=(Y-30)/4$	d_x^2	d_y^2	$d_x d_y$
1	46	−4	4	16	16	−16
2	42	−3	3	9	9	− 9
3	38	−2	2	4	4	− 4
4	34	−1	1	1	1	− 1
5	30	0	0	0	0	0
6	26	1	−1	1	1	− 1
7	12	2	−2	4	4	− 4
8	18	3	−3	9	9	− 9
9	14	4	−4	16	16	−16
10	10	5	−5	25	25	−25
Total		5 $(=\Sigma d_x)$	−5 $(=\Sigma d_y)$	85 $(=\Sigma d_x^2)$	85 $(=\Sigma d_y^2)$	−85 $(\Sigma d_x d_y)$

$$\gamma = \frac{n\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{\sqrt{n\Sigma d_x^2 - (nd_x)^2} \times \sqrt{n\Sigma d_y^2 - (\Sigma d_y)^2}}$$

$$= \frac{10(-85) - 5(-5)}{\sqrt{10 \cdot 85 - (5)^2} \times \sqrt{10 \cdot 85 - (-5)^2}}$$

$$= \frac{-850 + 25}{\sqrt{850 - 25} \times \sqrt{850 - 25}}$$

$$= \frac{-825}{\sqrt{825} \times \sqrt{825}} = \frac{-825}{825} = -1.$$

The variate X and Y are perfectly negatively correlated.

Example.

Calculate the Pearson's coefficient of correlation from the following data using 44 and 26 respectively as the origin of X and Y :

X :	43	44	46	40	44	42	45	42	38	40	42	57
Y :	29	31	19	18	19	27	27	29	41	30	26	10

[C.A. May '78]

X	Y	d_x	d_y	d_x^2	d_y^2	d_xd_y
		$=X-44$	$=Y-26$			
43	29	−1	3	1	9	−3
44	31	0	5	0	25	0
46	19	2	−7	4	49	−14
40	18	−4	−8	16	64	32
44	19	0	−7	0	49	0
42	27	−2	1	4	1	− 2
45	27	1	1	1	1	1
42	29	−2	3	4	9	− 6
38	41	−6	15	36	225	−90
40	30	−4	4	16	16	−16
42	26	−2	0	4	0	0
57	10	13	−16	169	256	−208
Total		−5	−6	255	704	−306
		$(=\Sigma d_x)$	$(=\Sigma d_y)$	$(=\Sigma d_y^2)$	$(=\Sigma d_x^2)$	$(=\Sigma d_xd_y)$

$$\gamma=\frac{n\Sigma d_xd_y-\Sigma d_x\Sigma d_y}{\sqrt{n\Sigma d_x^2-(\Sigma d_x)^2}\times\sqrt{n\Sigma d_y^2-(\Sigma d_y)^2}}$$

$$=\frac{12(-306)-(-5)(-6)}{\sqrt{12.255-(-5)^2}\times\sqrt{12.704-(-6)^2}}$$

$$=\frac{-3672-30}{\sqrt{3060-25}\times\sqrt{8448-36}}=\frac{-3702}{\sqrt{3035}\times\sqrt{8412}}.$$

Let $\gamma=\dfrac{3702}{\sqrt{3035}\times\sqrt{8412}}$

or, $\log\gamma=\log 3702-\frac{1}{2}\log(3035)-\frac{1}{2}\log(8412)$

$=3\cdot5684-\frac{1}{2}(3\cdot4821)-\frac{1}{2}(3\cdot9249)$

$=3\cdot5684-1\cdot7411-1\cdot9625=3\cdot5684-3\cdot7036$

$=-\cdot1352=-1+1-\cdot1352=-1+\cdot8648$

$=\bar{1}\cdot8648$

∴ $\gamma=\text{antilog }\bar{1}\cdot8648=\cdot7324$

∴ $\gamma=-0\cdot7324=-0\cdot732.$

Properties of γ. [C. U. B. Com. (Hons.) 1980]

(A) Correlation coefficient is a pure number, *i.e.*, it is independent to the unit of measurement if variable.

(B) The correlation coefficient does not depend on origin of reference or scale of measurement.

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be a set of n pairs of observations and also

$u_i = \dfrac{x_i - A}{C}$ and $v_i = \dfrac{y_i - B}{D}$ where A, B, C and D are four arbitrary constants.

Then $x_i = A + Cu_i$ and $y_i = D + Bv_i$

$\therefore \quad \bar{x} = A + C\bar{u}$ and $\bar{y} = D + B\bar{v}$.

Also $(x_i - \bar{x}) = A + Cu_i - (A + C\bar{u}) = C(u_i - \bar{u})$

and $(y_i - \bar{y}) = D + Bv_i - (D + B\bar{v}) = D(v_i - \bar{v})$.

Also $\text{var }(x) = \dfrac{1}{n} \Sigma(x_i - \bar{x})^2 = \dfrac{1}{n} \Sigma C^2 (u_i - \bar{u})^2$

$$= C^2 \cdot \frac{1}{n} \Sigma(u_i - \bar{u})^2 = C^2 \text{ var } (u)$$

$\therefore \quad \sigma_x = |C| \sigma_u$.

Similarly $\sigma_y = |D| \sigma_v$.

Again, $\text{cov }(x, y) = \dfrac{1}{n} \Sigma(x_i - \bar{x})(y_i - \bar{y}) = \dfrac{1}{n} \Sigma C (u_i - \bar{u}).D (v_i - \bar{v})$

$$= C.D. \frac{1}{n} \Sigma(u_i - \bar{u})(v_i - \bar{v}) = C.D. \text{ cov } (u, v)$$

$$\therefore \quad \gamma_{xy} = \frac{\text{cov }(x, y)}{\sigma_x . \sigma_y} = \frac{C.D.\text{cov }(u, v)}{|C|\,|D|\,\sigma_u \sigma_v} = \frac{C.D}{|C|\,|D|} \cdot \gamma_{uv}.$$

Case A. When C and D are of same sign, then C.D is positive and $\dfrac{C.D}{|C|\,|D|} = 1$ and hence γ_{xy} and γ_{uv} are equal in magnitude and sign.

Case B. When C and D are of different sign, then C.D is negative and $\dfrac{C.D}{|C|\,|D|} = -1$ and hence γ_{xy} and γ_{uv} are equal in magnitude but opposite in sign.

(C) The correlation coefficient lies between -1 and $+1$.

Let $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ be a set of n pairs of observations and

$$x_i' = \frac{x_i - \bar{x}}{\sigma_x} \quad \text{and} \quad y_i' = \frac{y_i - \bar{y}}{\sigma_y}.$$

Then $\Sigma x_i^2 = \sum \frac{(x_i - \bar{x})^2}{\sigma_x^2} = \frac{1}{\sigma_x^2} n\sigma_x^2 = n.$

Similarly $\Sigma y_i^2 = n$

$$\text{and} \quad \gamma = \frac{\frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y})}{\sigma_x \sigma_y} = \frac{1}{n}\Sigma x_i' y_i'.$$

Since the sum of squares of real numbers cannot be negative, we have

$$\frac{1}{n}\Sigma(x_i' + y_i')^2 > 0$$

or, $\frac{1}{n}\Sigma x_i'^2 + \frac{1}{n}\Sigma y_i'^2 + \frac{1}{n}\Sigma 2x_i'y_i' > 0$

or, $\frac{1}{n}\cdot n + \frac{1}{n}\cdot n + 2\gamma > 0$

or, $2 + 2\gamma > 0$

or, $2(1+\gamma) > 0$

or, $1 + \gamma > 0$

or, $\gamma > -1.$

Again $\frac{1}{n}\Sigma(x_i' - y_i')^2 > 0$

or, $\frac{1}{n}\Sigma x_i'^2 + \frac{1}{n}\Sigma y_i'^2 - \frac{2}{n}\Sigma x_i'y_i' > 0$

or, $2 - 2\gamma > 0$

or, $2(1-\gamma) > 0$

or, $1 - \gamma > 0$

or, $\gamma < 1.$

Thus the correlation coefficient must lie between -1 and $+1$.

(3) *Rank Method* (*Rank Correlation*).

There are some attributes (intelligence, honesty, character, morality, leadership, etc.) which cannot be measured by quantity. In such cases individuals in the group can be arranged in order and hence obtaining for each individual a number indicating the rank in the group.

Suppose the values of a variable (weight in kg.) are 50, 53, 54, 47, 59. If these figures are arranged in descending order, the figure 59 would receive the 1st rank, 54—2nd, 53—3rd, 50—4th, 49—5th rank. The rank of the variable whose value is highest is 1 and so on.

If again there are two or more items having the same value, then the process of distributing rank is as follows.

Let two items have equal value and their rank is 4. Now the two items will be given average rank of the ranks which they would get had there been slight difference in values. So the average rank would be $\frac{4+5}{2}=4{\cdot}5$ and the rank of the next item would be 6 (and not 5).

Note. Rank may be assigned either in ascending or in descending order.

Now the process of calculating the coefficient of correlation (γ) is as follows :

(i) Assign ranks to various items of the two series (if it is not given)

(ii) Find differences of the ranks (d)

(ii) Square these differences (d^2)

(iii) Use the following formula for finding (γ) :

$$\gamma=1-\frac{6(\Sigma d^2)}{n^3-n}, \text{ where } n=\text{number of pairs of observations.}$$

This method was developed by *C. E. Spearman*, British Psychologist, in 1904.

The value of this coefficient ranges between +1 and −1. If $\gamma=+1$, there is complete agreement in the order of ranks and the ranks are in the same direction. Again if $\gamma=-1$, there is complete agreement in the order of ranks and they are in opposite directions.

In rank correlation we have two types of problems :

(*a*) *Where actual ranks are given.*

(*b*) *Where ranks are not given.*

(*a*) WHERE ACTUAL RANKS ARE GIVEN.

Example.

Ten competitors in a voice contest are ranked by three judges in the following order :

1st Judge :	1	6	5	10	3	2	4	9	7	8
2nd Judge :	3	5	8	4	7	10	2	1	6	9
3rd Judge :	6	4	9	8	1	2	3	10	5	7

Use the method of rank-correlation to judge which pair of judges have the nearest approach to common liking in voice.

Ranks given by			*differences* *(d)*			*Squares of differences* (d^2)		
1st Judge	2nd Judge	3rd Judge	(i)	(ii)	(iii)	(i)	(ii)	(iii)
1	3	6	−2	−3	−5	4	9	25
6	5	4	1	1	2	1	1	4
5	8	9	−3	−1	−4	9	1	16
10	4	8	6	−4	2	36	16	4
3	7	1	−4	6	2	16	36	4
2	10	2	−8	8	0	64	64	0
4	2	3	2	−1	1	4	1	1
9	1	10	8	−9	−1	64	81	1
7	6	5	1	1	2	1	1	4
8	9	7	−1	2	1	1	4	1
						200	214	60

(i) $\gamma_{12} = 1 - \dfrac{6(\Sigma d^2)}{n^3 - n} = 1 - \dfrac{6.200}{10^3 - 10} = 1 - \dfrac{1200}{1000 - 10}$

(For 1st & 2nd judgment)

$$= 1 - \frac{1200}{990} = 1 - 1{\cdot}213 = -0{\cdot}213.$$

(ii) $\gamma_{23} = 1 - \dfrac{6(\Sigma d^2)}{n^3 - n} = 1 - \dfrac{6.214}{10^3 - 10} = 1 - \dfrac{1284}{990}$

(For 2nd & 3rd judgment)

$$= 1 - 1{\cdot}297 = -0{\cdot}297.$$

(iii) $\gamma_{13} = 1 - \dfrac{6(\Sigma d^2)}{n^3 - n} = 1 - \dfrac{6.60}{1000 - 10} = 1 - \dfrac{360}{990}$

(For 1st & 3rd judgment)

$$= 1 - {\cdot}364 = +0{\cdot}636.$$

The results of these coefficients indicate that the first and third judges have the nearest approach to common liking in voice.

(*b*) WHERE RANKS ARE NOT GIVEN.

Example.

The following are the marks obtained by 8 students in English and Bengali papers. Compute rank coefficient of correlation.

Marks in English :	15	20	28	12	40	60	20	80
Marks in Bengali :	40	30	50	30	20	10	30	60

Computation of Rank Correlation

Marks in English (X)	rank	Marks in Bengali (Y)	rank	difference d	d^2
15	2	40	6	−4	16
20	3·5	30	4	−·5	·25
28	5	50	7	−2	4
12	1	30	4	−3	9
40	6	20	2	4	16
60	7	10	1	6	36
20	3·5	30	4	−·5	·25
80	8	60	8	0	0
Total	—	—	—	—	81·50

For equal ranks some adjustment in the above formula is required, *i e.*, to add $\frac{1}{2}(m^3-m)$ with Σd^2 where m = number of items whose ranks are common.

Here, $$\gamma=1-\frac{6\{\Sigma d^2+\frac{1}{12}(m^3-m)+\frac{1}{12}(m^3-m)\}}{n^3-n}$$

The item 20 is repeated 2 times in X-series, *i.e.*, $m=2$ in X-series and again $m=3$ in Y-series.

$$\therefore\quad \gamma=1-\frac{6\{81{\cdot}5+\frac{1}{12}(2^3-2)+\frac{1}{12}(3^3-3)\}}{8^3-8}$$

$$=1-\frac{6\{81{\cdot}5+{\cdot}5+2\}}{504}=1-\frac{6.84}{504}=1-\frac{504}{504}=1-1=0.$$

There is no correlation.

Note. Some statisticians prefer the previous formula without adjustment.

Bivariate Data.

Previously the methods of summarisation of data having variation of one character have been discussed. Very often data may relate to variation in two or more characters. At present, let us take two variables, represented by x and y. Thus x may be the height and y weight of a person. The observations of each person (*i.e.*, individual) are paired. Thus for n observations we will find (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n). Statistical data relating to simultaneous measurement of two variables are called *Bivariate Data.* Again data relating to one variable only are called *Univariate Data.*

Example.

Bivariate data of height (x) in cms. and weight (y) in kgs. of 15 persons :

x :	174	170	178	175	180	172	168	174	180	170	161	172	171	177	181
y :	62	59	64	62	67	60	59	64	69	60	56	64	61	65	69

When the number of pairs of observations is large, then it is necessary to form a two-way frequency table, usually known as *Bivariate Frequency Table* or *Correlation Table.** The method of framing such table is similar to (univariate) frequency distributions table (discussed earlier). For the n pairs of observations (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n) in relation to two variables x and y (taken above), the class-intervals of both the series are chosen first based on extreme values of x and y series. If there are m class-intervals for x series and n class-intervals for y series, a table with m rows and n columns is to be constructed. So in the table, there will be $m \times n$ rectangular spaces, known as *cells.* The class-intervals of x and y series are taken as row-headings and column-headings respectively.

Procedure of framing bivariate frequency distribution of the data given above.

There are many different numbers both for x and y variates. In this case it is better to make some suitable class-intervals for x and y series. In this case for x-series, class-intervals 161—165, 166 − 170, $\cdots$, etc. and those for y-series, class-intervals 56—59, 60—63, $\cdots$, etc. are taken.

* A correlation table is also known as Bivariate Frequency Table, since it shows the frequency distribution of two related variables.

Now the first number of x series (174) lies in the class (171—175) and first number of y series (62) lies in the class (60—63). Now in the cell corresponding to (171—175) and (60—63) (*i.e.*, 2nd cell of 3rd row), one tally mark is given. Again for the second numbers of x and y series (*i.e.*, 170 and 59) lie in the classes (166—170) and (56—59) respectively. So in their corresponding cell (*i.e.*, the 1st cell of the 2nd row) one tally mark is given. Similarly for the rest numbers tally marks are given. Now the markings are to be counted and to be written in the column and row totals.

The bivariate frequency distribution of the above examples is shown below :

Bivariate Frequency Distribution

Weight (kg.)

	x / y	56—59	60—63	64—67	68—71	Total
height (cm.)	161—165	/ 1				1
	166—170	// 2	/ 1			3
	171—175		//// 4	// 2		6
	176—180			/// 3	/ 1	4
	181—185				/ 1	1
	Total	3	5	5	2	15

Here the number of observations lying in a cell is known as *cell frequency.*

Note. Here in x variates, there are 5 class-intervals and in y variates 4 class-intervals. So total number of cells is $5 \times 4 = 20$.

Class-intervals of x and y series may also be placed in column-headings and row-headings respectively.

Marginal Distribution and Conditional Distribution.

A univariate distribution (say x variable) obtained from the bivariate distribution, irrespective of values of other variable (y variable) is called a *marginal distribution.* The column totals of frequencies show the number of individuals belonging to x variate. This shows the frequency distribution of x, known as *marginal distribution of* x in the present context.

Example.

Marginal Distribution of Height

height (cm.)	frequency
161—165	1
166—170	3
171—175	6
176—180	4
181—185	1
Total	15

Marginal Distribution of Weight

weight (kg.)	frequency
56—59	3
60—63	5
64—67	5
68—71	2
Total	15

Again a univariate distribution obtained from a bivariate distribution for a particular value of the other variate is known as *conditional distribution.* Thus we find only one conditional distribution of x variate corresponding to each class-interval of y variate. Similarly, there is only one conditional distribution of y corresponding to each class-interval of x variate.

Example.

Conditional Distribution of Height
(when weight 64—67 kg.)

height (cm.)	frequency
161—165	0
166—170	0
171—175	2
176—180	3
181—185	0
Total	5

Conditional Distribution of Weight
(when height 166—170 cm.)

weight (kg.)	frequency
56—59	2
60—63	1
64—67	0
68—71	0
Total	3

Conditional Mean Value

The arithmetic mean of the specified conditional distribution is called the conditional mean value.

Example.

Find the conditional mean values of x for $y = 6, 10$ from the following bivariate frequency distribution :

x \ y	4	6	8	10	Total
5	2	1	1	3	7
10	1	2	1	1	5
15	0	1	1	1	3
Total	3	4	3	5	15

Calculation of Conditional Mean Values of x

(*a*) *when* $y = 6$

x	f	fx
5	1	5
10	2	20
15	1	15
Total	4	40

$$\text{A.M. } (\bar{x}) = \frac{\Sigma fx}{\Sigma f} = \frac{40}{4} = 10.$$

(*b*) *when* $y = 10$

x	f	fx
5	3	15
10	1	10
15	1	15
Total	5	40

$$\text{A.M. } (\bar{x}) = \frac{\Sigma fx}{\Sigma f} = \frac{40}{5} = 8.$$

Example.

The data given below relate to heights and weights of 20 persons. You are required to form a two-way frequency table with class-intervals 62" to 64", 64" to 66" and so on, and 115 to 125 lbs., 125 to 135 lbs. and so on.

Sl. no. :	1	2	3	4	5	6	7	8	9	10
Height :	70	65	65	64	69	63	65	70	71	62
Weight :	170	135	136	137	148	124	117	128	143	129

Sl. no. :	11	12	13	14	15	16	17	18	19	20
Height :	70	67	63	68	67	69	66	68	67	67
Weight :	163	139	122	134	140	132	120	148	129	152

[C. A. May 1966]

wt (*y*) / ht. (*x*)	115—125	125—135	135—145	145—155	155—165	165—175	Total
62—64	//	/					3
64—66	/		///				4
66—68	/	/	//	/			5
68—70		//		//			4
70—72		/	/		/	/	4
Total	4	5	6	3	1	1	20

Note. Here class-interval 62—64 (for height) means 62 and less than 64, and similarly for all other class-intervals of height and weight (*i.e.*, continuous series).

Here tally marks are used to indicate frequencies.

Calculation of Correlation Coefficient from Grouped Data

When the number of observations of X and Y variables is large, the data are classified into a correlation table. The formula used is as follows—

$$\gamma = \frac{n\Sigma fd_x d_y - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{n\Sigma fd_x^2 - (\Sigma fd_x)^2}\cdot\sqrt{n\Sigma fd_y^2 - (\Sigma fd_y)^2}}.$$

Note. This formula is same as discussed above, while deviations are taken from assumed mean. The only difference is that here the deviations are also multiplied by the frequencies.

Explanations of the Symbols Used in the following examples :

(i) x, y indicate mid-values of X and Y series respectively.

(ii) $d_x = \dfrac{x-140}{10}$ and $d_y = \dfrac{y-67}{2}$.

(iii) f_1, f_2 represent marginal frequencies of x and y distributions respectively.

		Y	115—125	125—135	135—145	145—155	155—165	165—175	totals			
		y	120	130	140	150	160	170				
		d_y	-2	-1	0	1	2	3	f_1	f_1d_x	$f_1d_x^2$	fd_xd_y
X	x	d_x										
62—64	63	-2	2 (8)*	1 (2)					3	-6	12	10
64—66	65	-1	1 (2)		3 (0)				4	-4	4	2
66—68	67	0	1 (0)	1 (0)	2 (0)	1 (0)			5	0	0	0
68—70	69	1		2 (−2)		2 (2)			4	4	4	0
70—72	71	2		1 (−2)	1 (0)		1 (4)	1 (6)	4	8	16	8
totals	f_2		4	5	6	3	1	1	$n=20$	2	36	20
	f_2d_y		-8	-5	0	3	2	3	-5			
	$f_2d_y^2$		16	5	0	3	4	9	37			
	fd_xd_y		10	-2	0	2	4	6	20			

* (8) = 2 × (−2) × (−2)

(iv) number within brackets in every cell is the product of that cell frequency and the corresponding values of d_x and d_y.

(v) fd_xd_y is the total of the numbers within brackets mentioned in (iv), in any row or column.

So the above formula may be written as follows (for convenience):

$$\gamma=\frac{n\Sigma fd_xd_y-(\Sigma f_1d_x)(\Sigma f_2d_y)}{\sqrt{n\Sigma f_1d_x^2-(\Sigma f_1d_x)^2}\,.\,\sqrt{n\Sigma f_2d_y^2-(\Sigma f_2d_y)^2}}.$$

Example.

Calculation of coefficient of correlation of the above example is shown in the previous page (p. 266).

$$\text{Now,}\quad \gamma=\frac{n\Sigma fd_xd_y-(\Sigma f_1d_x)(\Sigma f_2d_y)}{\sqrt{n\Sigma f_1d_x^2-(\Sigma f_1d_x)^2}\,.\,\sqrt{n\Sigma f_2d_y^2-(\Sigma f_2d_y)^2}}$$

$$=\frac{20.20-2.(-5)}{\sqrt{20.36-2^2}\,.\,\sqrt{20.37-(-5)^2}}=\frac{400+10}{\sqrt{720-4}\,.\,\sqrt{740+25}}$$

$$=\frac{410}{\sqrt{716}\,.\,\sqrt{765}}$$

$$\text{or, }\log\gamma=\log 410-\tfrac{1}{2}(\log 716+\log 765)$$

$$=2{\cdot}6128-\tfrac{1}{2}(2{\cdot}8549-2{\cdot}8837)$$

$$=2{\cdot}6128-\tfrac{1}{2}(5{\cdot}7386)$$

$$=2{\cdot}6128-2{\cdot}8693$$

$$=-0{\cdot}2565$$

$$=\bar{1}{\cdot}7435$$

$$\therefore\quad \gamma=\text{antilog }\bar{1}{\cdot}7435={\cdot}554.$$

EXERCISE 9

1. (*a*) Define simple correlation. What is the difference between a positive correlation and negative correlation ?

[I.C.W.A. June '75, June '77, C.A. Nov. '75]

(*b*) Define correlation coefficient and state its important properties (clearly explain all the symbols you use).

[C.U. B.Com. (Hons.) 1980]

2. Mention few methods of studying correlation. What is Scatter Diagram. Indicate by means of suitable scatter diagram different types of correlation that may exist between the variables in bivariate data. [I.C.W.A. June '74, 7une '76]

3. Write notes on :
   (i) Scatter Diagram.
   (ii) Pearsonian Coefficient of Correlation.
   (iii) Rank Correlation Coefficient.

4. Following are the heights and weights of 10 students of a B.Com. class :

Height (inch) :	62	72	68	58	65	70	66	63	60	72
Weight (kg.) :	50	65	63	50	54	60	61	55	54	65

Draw a scatter diagram and indicate whether the correlation is positive or negative.

5. Construct a scatter diagram of the data given below and fit a straight line by free-hand method :

(*Average value in lakh Rs.*)

Years :	1965	1966	1967	1968	1969	1970	1971	1972
Export :	47	64	100	97	126	203	171	115
Import :	70	85	100	103	111	139	133	115

(of a commodity).

6. Calculate Pearson's Coefficient of Correlation between Advertisement cost and sales as per the data given below :

Advertisement cost in '000 Rs. :	39	65	62	90	82	75	25	98	36	78
sales in lakh Rs. :	47	53	58	86	62	68	60	91	51	84

[C.A. Nov. 75] (*Ans.* 10·78)

7. Making use of the data given below, calculate the coefficient of correlation :

Case	X	Y	*Case*	X	Y
1	10	9	5	12	11
2	6	4	6	13	13
3	9	6	7	11	8
4	10	9	8	9	4

(*Ans.* +0·896)

8. Nine students obtained the following percentage of marks in

College Test (X) and in Final University Examination (Y). Calculate the correlation coefficient.

X :	51	63	73	46	50	60	47	36	60
Y :	49	72	74	44	58	66	50	30	35

[I.C.W.A. 1967] (*Ans.* +0·932)

9. Calculate Pearson's Coefficient of Correlation from the following, taking 100 and 50 as the assumed average of X and Y respectively :

X :	104	111	104	114	118	117	105	108	106	100	104	105
Y :	57	55	47	45	45	50	64	63	66	62	69	61

[C.A. Nov. 1976] (*Ans.* −0·67)

10. Find the correlation coefficient between the income and expenditure of a wage-earner and comment :

Month :	*Jan.*	*Feb.*	*March*	*April*	*May*	*June*	*July*
Income :	46	54	56	56	58	60	62
Expenditure :	36	40	44	54	42	58	54

(*Ans.* +0·769)

11. Calculate the coefficient of correlation for the ages of husband and wife :

Age of husband :	23	27	28	29	30	31	33	35	36	39
Age of wife :	18	22	23	24	25	26	28	29	30	32

[I.C.W.A. 1970] (*Ans.* +0·995)

12. Marks in Mathematics and Statistics of 10 students are given below :

Math. :	32	38	48	43	40	22	41	69	35	64
Stat. :	30	31	38	43	33	11	27	76	40	59

—Find the coefficient of correlation.

[I.C.W.A. June '74] (*Ans.* +0·935)

13. The table below gives the respective heights X and Y of two samples of 10 plants each grown under two different conditions :

Plant No.	*Sample I X (cm.)*	*Sample II Y (cm.)*
1	30	45
2	50	63
3	42	55
4	25	48
5	60	65
6	28	48
7	32	50
8	55	60
9	58	60
10	35	49

—Find the value of γ. (*Ans.* +0·9)

14. Calculate rank correlation coefficient for a group of 10 students in the following case :

Marks in History :	70	65	63	60	58	55	54	53	50	45
Marks in Geography :	90	80	76	75	70	50	48	45	42	40

(*Ans.* 1)

15. Following are the ranks obtained by 10 students in two subjects—Statistics and Mathematics. To what extent the knowledge of students in the two subjects is related ?

Statistics :	1	2	3	4	5	6	7	8	9	10
Mathematics :	2	4	1	5	3	9	7	10	6	5

(*Ans.* +0·76)

16. Calculate coefficient of rank correlation from the following data :

x :	48	33	40	9	16	16	65	24	16	57
y :	13	13	24	6	15	4	20	9	9	19

(*Ans.* +0·733)

17. From the following data calculate the coefficient of rank correlation between x and y :

x :	36	56	20	65	42	33	44	50	15	60
y :	50	35	70	25	58	75	60	45	80	38

(*Ans.* −0·927)

18. In a contest, two judges ranked eight candidates A, B, C, D, E, F, G and H in order of their performance, as shown in the following table. Find the rank correlation coefficient :

	A	B	C	D	E	F	G	H
First judge :	5	2	8	1	4	6	3	7
Second judge :	4	5	7	3	2	8	1	6

[I.C.W.A. June '75] (*Ans.* $+\frac{2}{3}$)

19. Find the value of coefficient of correlation from the following values :

(*a*) $\sum_{i=1}^{11}(X_i-\bar{X})^2=110\,;\ \sum_{i=1}^{11}(Y_i-\bar{Y})^2=990\,;\ \sum_{i=1}^{11}(X_i-\bar{X})(Y_i-\bar{Y})=330$

(*Ans.* 1)

(*b*) $\sum_{i=1}^{100} X_i=280\,;\ \sum_{i=0}^{100} Y_i=60\,;\ \sum_{i=1}^{100} X_i^2=2{,}384\,;\ \sum_{i=1}^{100} Y_i^2=117\,;$

$\sum_{i=1}^{100} X_iY_i=438.$

[I.C.W.A. July '71] (*Ans.* 0·75)

20. Marks obtained in Statistics and Auditing by 24 students are given below. Prepare a bivariate frequency distribution table.

Sl. No.	*Marks in Stat.*	*Marks in Audit.*	*Sl. No.*	*Marks in Stat.*	*Marks in Audit.*	*Sl. No.*	*Marks in Stat.*	*Marks in Audit.*
1	22	16	9	22	16	17	27	15
2	23	16	10	23	18	18	27	16
3	23	18	11	24	18	19	26	18
4	23	16	12	24	17	20	28	19
5	23	16	13	23	16	21	25	19
6	24	17	14	25	17	22	24	16
7	23	16	15	23	17	23	23	17
8	25	19	16	22	17	24	25	19

Hints : The marks in Statistics assume only 7 values from 22—28 and those of Auditing assume only 5 values from 15—19. Against these values tally marks are to be plotted to form the table, similar to discrete series table. (*Ans.* freq. : 3, 9, 4, 4, 1, 2, 1
freq. : 1, 9, 6, 4, 4)

21. The ages of 20 husbands and wives are given below. Form a two-way frequency table showing the relationship between the ages of husbands and wives with the class-intervals 20—25, 25—30, etc.
[C. A. May, 1970]

Sl. No. :	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Age of husband :	28	37	42	25	29	47	37	35	23	41	27	39	23	33	36	32	22	29	38	48
Age of wife :	23	30	40	26	25	41	35	25	21	38	24	34	20	31	29	35	23	27	34	47

(*Ans.* freq. : 5, 5, 4, 3, 2, 1 ; freq. : 3, 5, 2, 6, 2, 2)

22. From the following table obtain the conditional mean values of y for given values of $x=0, 1, 2, 3$:

$y \backslash x$	0	1	2	3	4
0	0	0	0	4	3
1	0	0	18	36	9
2	0	12	54	36	3
3	1	12	18	4	0

(*Ans.* 3, 2·5, 2, 1·5)

23. Calculate the coefficient of correlation between the marks obtained by a batch of 100 students in Accountancy and Statistics as given in the following table :

Marks in Statistics	*Marks in Accountancy*					
	20—30	30—40	40—50	50—60	60—70	Total
15—25	5	9	3			17
25—35		10	25	2		37
35—45		1	12	2		15
45—55			4	16	5	25
55—65				4	2	6
Total	5	20	44	24	7	100

(*Ans.* 0·795)

24. The following are the marks obtained by the students of a class in Statistics and Accountancy :

Sl. No.	*Marks in Statistics*	*Marks in Account.*	*Sl. No.*	*Marks in Statistics*	*Marks in Account.*
1	15	13	13	14	11
2	0	1	14	9	3
3	1	2	15	8	5
4	3	7	16	13	4
5	16	8	17	10	10
6	2	9	18	13	11
7	18	12	19	11	14
8	5	9	20	11	7
9	4	17	21	12	18
10	17	16	22	8	15
11	6	6	23	9	15
12	19	18	24	7	3

Prepare a correlation table taking the magnitude of each class-interval as 4 marks and the first class interval as equal to 0 and less than 4. Calculate Karl Pearson's coefficient of correlation between the marks in Statistics and Accountancy. (*Ans.* 0·578)

25. The following table gives the frequency according to age-groups of marks obtained by 67 students in an intelligence test. Measure the degree of relationship between age and general knowledge.

Age in Years.

Test marks	18	19	20	21	*Total*
200—250	4	4	2	1	11
250—300	3	5	4	2	14
300—350	2	6	8	5	21
350—400	1	4	6	10	21
Total	10	19	20	18	67

(*Ans.* 0·415)

26. Family income and its percentage spent on food in the case of hundred families gave the following bivariate frequency distribution. Calculate the coefficient of correlation.

Family Income in Rupees

Food expenditure :	200—300	300—400	400—500	500—600	600—700
10—15%	—	—	—	3	7
15—20%	—	4	9	4	3
20—25%	7	6	12	5	—
25—30%	3	10	19	8	—

(*Ans.* 0.438)

27. The following table gives a bivariate frequency distribution of 50 clerks according to age in years and pay in rupees ; find the value of the correlation coefficient between the variables.

Pay

Age	250—300	300—350	350—400	400—450	*Total*
20—30	8	3	—	—	11
30—40	2	5	2	2	11
40—50	—	2	9	6	17
50—60	—	—	5	6	11
Total	10	10	16	14	50

(*Ans.* 0·76)

(B) Regression.

Introduction : In the previous chapter, we have established the close relation between two variables. Now we are interested to estimate (predict) the value of one variable for the given value of the other. For example, the heights and weights are correlated, now we may estimate a height for a given weight.

The term 'regression' means, going back or study, according to dictionary. F. Galton's study regarding the height of fathers and their sons revealed an interesting relationship. The deviations of mean heights of the sons from the mean height of the race were less than the deviations in the mean height of the fathers from mean height of the race. When the fathers were above or below the mean, the sons tended to go back or *regress* towards the mean. Thus regression implies going back or returning. Galton represented the average relationship between these variables graphically and called the line thus obtained as the *line of regression.* Regression lines give idea on the correlation of two series. If the coefficient or correlation between the heights of fathers and their sons is + ·6, it means of a group of fathers had an average of x cms. above general average, the average height of their sons would be only + ·6x cms. above the general average. This going back towards the average is called *regression.*

The *regression analysis* helps in following ways :—

(1) To estimate (or predict) the values of dependent variables from values of independent variable.

(2) To obtain the measure of error involved in using the regression line as a basis of estimation.

(3) To obtain a measure of association or correlation that exists between the two variables.

At present we shall do problems of two variables (*i.e.*, simple regression), although the analysis may be extended to three or more variables.

Difference between Correlation and Regression.

Correlation coefficient is a measure of *degree of relationship* between X and Y, whereas the regression analysis reveals the study of *nature of relationship* between the variables.

Regression Equations.

For two variables x and y, we shall have two regression lines. One regression of x on y and the other of y on x. Regression equations are algebraic expression of regression lines. Now for two regression lines there will be two regression equations. It may be noted that the regression of x on y is used to describe the variation in the values of x for given changes in y and the regression equation of y on x is used to describe the variation in the values of y for given changes in x.

Regression Equation of X on Y :

This equation is as follows,

$$x = a + by.$$

Now to determine the constants a and b we are to solve the following normal equations :

$\Sigma x = na + b\Sigma y$, n = number of observed pair of values ;
$\Sigma xy = a\Sigma y + b\Sigma y^2$.

Regression Equation of Y on X :

The equation is $y = a + bx$, where the value of a and b are to be obtained by solving :

$$\Sigma y = na + b\Sigma x$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2.$$

Example.

From the following data obtain the two regression equations :

x :	6	2	10	4	8
y :	9	11	5	8	7

[I. C. W. A. Jan. 1967]

Computation of Regression Equations

x	y	xy	x^2	y^2
6	9	54	36	81
2	11	22	4	121
10	5	50	100	25
4	8	32	16	64
8	7	56	64	49
$\Sigma x = 30$	$\Sigma y = 40$	$\Sigma xy = 214$	$\Sigma x^2 = 220$	$\Sigma y^2 = 340$

For equation of x on y : $x = a + by$ and the normal equations are

$$\Sigma x = na + b\Sigma y \quad \text{and} \quad \Sigma xy = a\Sigma y + b\Sigma y^2.$$

Now putting values, $30=5a+40b$ ... (i)

$214=40a+340b$ ... (ii)

Multiplying (i) by 8 and subtracting from (ii) we get,

$$-26=20b \quad \text{or,} \quad b=-1{\cdot}3.$$

Now putting this value of b in (i) we get

$$30=5a+40(-1{\cdot}3) \quad \text{or,} \quad a=16{\cdot}4.$$

∴ the regression line of x on y is $x=16{\cdot}4-1{\cdot}3y$.

Now for the other regression equation of y on x, we get

$$y=a+bx$$

and the normal equations as $\Sigma y=na+b\Sigma x$

$\Sigma xy=a\Sigma x+b\Sigma x^2$.

Putting values, $40=5a+30b$ ... (i)

and $214=30a+220b$ ... (ii)

Multiplying (i) by 6 and subtracting it from (ii) we get

$$-26=40b \quad \text{or,} \quad b=-0{\cdot}65.$$

Putting this value of b in (i), we have

$$40=5a+30(-{\cdot}65)$$

or, $a=11{\cdot}9$.

Hence the regression line of y on x is

$$y=11{\cdot}9-0{\cdot}65x.$$

To find the Regression Equation of y on x.

The regression equation of y on x can be represented in the form $y=a+bx$, where a and b are constants, and determine the position of the regression line completely.

Let $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ be a set of n observations of the two variates x and y. Through the set of n observations a straight line of the form

$$y=a+bx \quad \cdots \quad (1)$$

can be fitted. In this case, x is independent and y is dependent.

Now, to find the values of a and b, we are to apply the method of least square and solve the following normal equations :

$$\Sigma y \;=na+b\Sigma x \quad \cdots \quad (2)$$

$$\Sigma xy=a\Sigma x+b\Sigma x^2 \quad \cdots \quad (3)$$

In eqn. (2), dividing both sides by n, we get

$$\frac{\Sigma y}{n}=a+b\,\frac{\Sigma x}{n} \quad \text{or,} \quad \bar{y}=a+b\bar{x}. \qquad \cdots \quad (4)$$

Now, subtracting (4) from (1),

$$y-\bar{y}=b(x-\bar{x}) \qquad (5)$$

Again multiplying (2) by Σx and (3) by n, we find

$$(\Sigma x)(\Sigma y)=na(\Sigma x)+b(\Sigma x)^2$$
$$n(\Sigma xy)=na(\Sigma x)+nb(\Sigma x^2)$$

Subtracting, $(\Sigma x)(\Sigma y)-n(\Sigma xy)=b[(\Sigma x)^2-n(\Sigma x^2)]$

$$\text{or,}\quad b=\frac{n(\Sigma xy)-(\Sigma x)(\Sigma y)}{n(\Sigma x^2)-(\Sigma x)^2}, \text{ (changing sign)} \quad \cdots \quad (6)$$

$$\text{or,}\quad b=\frac{\dfrac{\Sigma xy}{n}-\dfrac{\Sigma x}{n}.\dfrac{\Sigma y}{n}}{\dfrac{\Sigma x^2}{n}-\left(\dfrac{\Sigma x}{n}\right)^2}, \text{ (dividing num. and deno. by } n^2\text{)}$$

$$=\frac{\text{cov}\,(x,\,y)}{\sigma_x^{\,2}}. \qquad \cdots \quad (7)$$

From (5), we find $y-\bar{y}=b_{yx}(x-\bar{x})$, where $b_{yx}=\dfrac{\text{cov}\,(x,\,y)}{\sigma_x^{\,2}}$

(b_{yx} indicates y on x).

This is the required regression equation of y on x.

Again we know $\gamma=\dfrac{\text{cov}\,(x,\,y)}{\sigma_x\cdot\sigma_y}$, *i.e.*, $\text{cov}\,(x,\,y)=\gamma\,\sigma_x\cdot\sigma_y$.

$$\text{From (6)},\; b_{yx}=\frac{\text{cov}\,(x,\,y)}{\sigma_x^{\,2}}=\gamma\,\frac{\sigma_y}{\sigma_x}. \qquad \cdots \quad (8)$$

Regression Equation of x on y.

Proceeding as above, the regression equation of x on y will be $x-\bar{x}=b(y-\bar{y})$, here b stands for b_{xy}

$$\text{where}\qquad b_{xy}=\frac{\dfrac{\Sigma xy}{n}-\dfrac{\Sigma x}{n}.\dfrac{\Sigma y}{n}}{\dfrac{\Sigma y^2}{n}-\left(\dfrac{\Sigma y}{n}\right)^2}, \qquad \cdots \quad (9)$$

$$=\frac{\text{cov}\,(x,\,y)}{\sigma_y^{\,2}}=\gamma\,\frac{\sigma_x}{\sigma_y}. \qquad \cdots \quad (10)$$

Example.

Find both the regression equations by the method mentioned in the previous example.

First, for regression equation of y on x, we have $y-\bar{y}=b_{yx}(x-\bar{x})$,

where $$b_{yx}=\frac{\text{cov}(x,y)}{\sigma_x^2}=\frac{\frac{\Sigma xy}{n}-\frac{\Sigma x}{n}.\frac{\Sigma y}{n}}{\frac{\Sigma x^2}{n}-\left(\frac{\Sigma x}{n}\right)^2}$$

$$=\frac{\frac{214}{5}-\frac{30}{5}.\frac{40}{5}}{\frac{220}{5}-\left(\frac{30}{5}\right)^2}=\frac{42\cdot8-6\times8}{44-(6)^2}$$

$$=\frac{42\cdot8-48}{44-36}=\frac{-5\cdot2}{8}=-0\cdot65.$$

Again $\bar{y}=\frac{\Sigma y}{n}=\frac{40}{5}=8$;

$\bar{x}=\frac{\Sigma x}{n}=\frac{30}{5}=6.$

$\therefore$ the regression equation of y on x is $y-8=-0\cdot65\,(x-6)$

or, $y-8=-0\cdot65x+3\cdot90$ *i.e.*, $y+0\cdot65x=11\cdot9$

(the same result as before)

Next for regression equation of x on y, we get $x-\bar{x}=b_{xy}(y-\bar{y})$.

Here, $$b_{xy}=\frac{\frac{\Sigma xy}{n}-\frac{\Sigma x}{n}.\frac{\Sigma y}{n}}{\frac{\Sigma y^2}{n}-\left(\frac{\Sigma y}{n}\right)^2}=\frac{\frac{214}{5}-\frac{30}{5}\times\frac{40}{5}}{\frac{340}{5}-\left(\frac{40}{5}\right)^2}=\frac{42\cdot8-6\times8}{68-8^2}$$

$$=\frac{42\cdot8-48}{68-64}=\frac{-5\cdot2}{4}=-1\cdot3.$$

$\therefore$ the regression equation of x on y is $x-6=-1\cdot3(y-8)$

or, $x+1\cdot3y=16\cdot4$ (the same result as before)

Properties of Linear Regression Equations.

1. The linear regression equation of y on x is $y-\bar{y}=b_{yx}(x-\bar{x})$ and that of x on y is $x-\bar{x}=b_{xy}(y-\bar{y})$ where b_{yx} and b_{xy} are known as regression coefficients of y on x and x on y respectively.

$$b_{yx}=\frac{\text{cov}(x,y)}{\sigma_x^2}=\gamma\,\frac{\sigma_y}{\sigma_x}\text{ and }b_{xy}=\frac{\text{cov}(x,y)}{\sigma_y^2}=\gamma\,\frac{\sigma_x}{\sigma_y}.$$

2. The product of two regression coefficients is equal to the square of the correlation coefficient,

$$i.e.,\ b_{yx} \times b_{xy} = \gamma\frac{\sigma_y}{\sigma_x} \times \gamma\frac{\sigma_x}{\sigma_y} = \gamma^2.$$

3. Regression coefficients and correlation coefficient, *i.e.*, b_{yx}, b_{xy} and γ have the same sign, *i.e.*, if both the regression coefficients have a negative sign, γ will also be negative, and again if the regression coefficients are both positive, then γ will be positive. If γ is zero, then b_{yx} and b_{xy} will be zero (it is clear from Prop. 2).

4. Two regression lines always intersect at $(\bar{x}, \bar{y})$. The slope of regression line of y on x is b_{yx} and that of x on y is $\frac{1}{b_{xy}}$.

5. Two regression equations may be written as $\frac{y-\bar{y}}{\sigma_y} = \gamma\frac{x-\bar{x}}{\sigma_x}$ and $\frac{x-\bar{x}}{\sigma_x} = \gamma\frac{y-\bar{y}}{\sigma_y}$ which are different. But if $\gamma = \pm 1$, the two equations become identical. Again if $\gamma = 0$, then we find $y = \bar{y}$ and $x = \bar{x}$. In that case y or x cannot be estimated from linear regression equations.

Show that the Correlation Coefficient is the Geometric Mean of Regression Coefficients.

We know $b_{yx} = \frac{\text{cov}(x, y)}{\sigma_x^2} = \gamma\frac{\sigma_y}{\sigma_x}$ ... (1)

(where γ = correlation coefficient)

and $b_{xy} = \frac{\text{cov}(x, y)}{\sigma_y^2} = \gamma\frac{\sigma_x}{\sigma_y}$ ... (2)

Now $b_{yx} \times b_{xy} = \gamma\frac{\sigma_y}{\sigma_x} \times \gamma\frac{\sigma_x}{\sigma_y} = \gamma^2$ [multiplying (1) and (2)]

or, $\gamma = \sqrt{b_{yx} \times b_{xy}}$ (a linear relation between correlation and regression coefficients)

Note. 1. We know $\gamma \leq 1$, *i.e.*, $\gamma \ngtr 1$, so also the product of regression coefficients cannot be greater than 1.

2. One of the regression coefficients must be less than or equal to 1.

3. Since γ is G. M. of two regression coefficients, the value of γ will lie between the values of two regression coefficients.

Show that regression coefficients are independent of change of origin but dependent on scale.

We know $b_{yx} = \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$ [see formula (6)]

$= \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$ [by (7)]

Let $u=\dfrac{x-A}{h}$ and $v=\dfrac{y-B}{k}$.

Then $x=A+hu$ and $y=B+kv$

or, $\bar{x}=A+h\bar{u}$ and $\bar{y}=B+k\bar{v}$.

Subtracting $x-\bar{x}=h(u-\bar{u})$ and $y-\bar{y}=k(v-\bar{v})$.

Putting these values in the above formula we get

$$b_{yx}=\frac{\Sigma hk(u-\bar{u})(v-\bar{v})}{\Sigma h^2(u-\bar{u})^2}=\frac{hk}{h^2}\cdot\frac{\Sigma(u-\bar{u})(v-\bar{v})}{\Sigma(u-\bar{u})^2}$$

$$=\frac{k}{h}\cdot\frac{\Sigma(u-\bar{u})(v-\bar{v})}{\Sigma(u-\bar{u})^2}=\frac{k}{h}b_{vu}.$$

Similarly we have, $b_{xy}=\dfrac{h}{k}b_{uv}$. Hence the result.

Deviations taken from actual A. M. of x and y.

We know the regression equation of y on x is

$y=a+bx$, here $b(=b_{yx})$ is the regression coefficient of y on x,

or $\bar{y}=a+b\bar{x}$.

Subtracting, $y-\bar{y}=b(x-\bar{x})$ which is the required regression equation of y on x.

Again $b(=b_{yx})=\dfrac{\text{cov }(x, y)}{\sigma_x^2}$ [by (7)]

$$=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$

$$=\frac{\Sigma XY}{\Sigma X^2}, \text{ where } X=x-\bar{x}\,;\ Y=y-\bar{y}.$$

Similarly $b_{xy}=\dfrac{\Sigma XY}{\Sigma Y^2}$.

∴ Regression equation of x on y is

$$x-\bar{x}=b_{xy}\,(y-\bar{y}).$$

Note. This formula is applicable when the actual mean is not in fraction.

Example.

Otain the equations of the two lines of regression for the data given below :

x :	1	2	3	4	5	6	7	8	9
y :	9	8	10	12	11	13	14	16	15

[I. C. W. A. Dec. 1978]

Calculation of Regression Lines

x	X $x-\bar{x}$	y	Y $y-\bar{y}$	X²	Y²	XY
1	−4	9	−3	16	9	12
2	−3	8	−4	9	16	12
3	−2	10	−2	4	4	4
4	−1	12	0	1	0	0
5	0	11	−1	0	1	0
6	1	13	1	1	1	1
7	2	14	2	4	4	4
8	3	16	4	9	16	12
9	4	15	3	16	9	12
45	0	108	0	60	60	57

$$\bar{x}=\frac{\Sigma X}{n}=\frac{45}{9}=5\,;\quad \bar{y}=\frac{\Sigma Y}{n}=\frac{108}{9}=12.$$

Regression coefficient of y on x, $b_{yx}=\dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}=\dfrac{\Sigma XY}{\Sigma X^2}$

$$=\frac{57}{60}=0{\cdot}95.$$

Regression line : $y-\bar{y}=b_{yx}\,(x-\bar{x})$

or $y-12=0{\cdot}95(x-5)$

or $y=0{\cdot}95x+7{\cdot}25.$

Again for regression coefficient of x on y,

$$b_{xy}=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2}=\frac{\Sigma XY}{\Sigma Y^2}=\frac{57}{60}=0{\cdot}95$$

∴ Regression line : $x-\bar{x}=b_{xy}\,(y-\bar{y})$

or, $x-5=0{\cdot}95(y-12)$

or, $x=0{\cdot}95y-6{\cdot}4.$

Example.

A department score gives in service training to its salesmen which is followed by a test. It is considering whether it should terminate the services of any salesman who does not do well in the test. The following data give the test scores and sales made by nine salesmen during a certain period :

Tests scores :	14	19	24	21	26	22	15	20	19
Sales ('00 Rs.) :	31	36	48	37	50	45	33	41	39

Calculate the coefficient of correlation between the test scores and the sales. Does it indicate that the termination of services of low cost scores is justified ? If the firm wants a minimum sales volume of Rs. 3,000, what is the minimum test score that will ensure continuation of service ? [C. A. Inter Nov. '74]

Let x denotes the test scores and y denotes the sales ('00 Rs.).

Calculation of Regression Equations

x	X $(x-\bar{x})$	X^2	y	Y $(y-\bar{y})$	Y^2	XY
14	−6	36	31	−9	81	54
19	−1	1	36	−4	16	4
24	4	16	48	8	64	32
21	1	1	37	−3	9	−3
26	6	36	50	10	100	60
22	2	4	45	5	25	10
15	−5	25	33	−7	49	35
20	0	0	41	1	1	0
19	−1	1	39	−1	1	1
180	0	120	360	0	346	193

$$\bar{x}=\frac{\Sigma x}{n}=\frac{180}{9}=20\ ;\ \bar{y}=\frac{\Sigma y}{n}=\frac{360}{9}=40$$

$$\sigma_x=\sqrt{\frac{\Sigma d^2}{n}}=\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}=\sqrt{\frac{\Sigma X^2}{n}}=\sqrt{\frac{120}{9}}=3{\cdot}65$$

$$\sigma_y=\sqrt{\frac{\Sigma Y^2}{n}}=\sqrt{\frac{346}{9}}=6{\cdot}2$$

$$\gamma=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{n\sigma_x\sigma_y}=\frac{\Sigma XY}{n\sigma_x\sigma_y}=\frac{193}{9\times3{\cdot}65\times6{\cdot}2}={\cdot}9477.$$

We find the coefficient of correlation is high enough to justify the proposal.

Now the regression of Test Scores (x) on sales (y) is given by

$$x-\bar{x}=b_{xy}(y-\bar{y}),$$

where $b_{xv}=\gamma\dfrac{\sigma_x}{\sigma_y}={\cdot}9477\times\dfrac{3{\cdot}65}{6{\cdot}2}$

$={\cdot}5578$ (calculation by log-table) or, $x-20={\cdot}5578\,(y-40)$

or, $x-20={\cdot}5578y-22{\cdot}3120$

or, $x={\cdot}5578y-2{\cdot}312$

For $y=$ Rs. 3000, *i.e.*, 30('00 Rs.) we have

$$x={\cdot}5578\times 30-2{\cdot}312$$
$$=16{\cdot}734-2{\cdot}312=14{\cdot}422\Rightarrow 14.$$

∴ A minimum test score 14 will ensure the continuation of service.

Example.

From the following results, obtain the two regression equations and estimate the yield of crops when the rainfall is 29 cms. and the rainfall when the yield is 600 kg.

	Y *(yield in kg.)*	X *(rainfall in cm.)*
Mean	508·4	26·7
S.D.	36·8	4·6

Coefficient of correlation between yield and rainfall $=0{\cdot}52$.

[C.A. Inter. May 1976]

Regression equation of y on x is

$$y-\bar{y}=b_{yx}\,(x-\bar{x}),\text{ where,}$$

$$b_{yx}=\gamma\frac{\sigma_y}{\sigma_x}=0{\cdot}52\times\frac{36{\cdot}8}{4{\cdot}6}=4{\cdot}16$$

or, $y-508{\cdot}4=4{\cdot}16\,(x-26{\cdot}7)$

or, $y-508{\cdot}4=4{\cdot}16x-111{\cdot}072$

or, $y=4{\cdot}16x+397{\cdot}328$

For $x=29$, we have $y=4{\cdot}16\times 29+397{\cdot}328$

$$=120{\cdot}64+397{\cdot}33=517{\cdot}97\text{ kg.}$$

Regression equation of x on y,

$$x-\bar{x}=b_{xy}\,(y-\bar{y})$$

where $b_{xy}=\gamma\dfrac{\sigma_x}{\sigma_y}=0{\cdot}52\times\dfrac{4{\cdot}6}{36{\cdot}8}={\cdot}065$

or, $x-26{\cdot}7={\cdot}065(y-508{\cdot}4)$

or, $x-26{\cdot}7={\cdot}065y-33{\cdot}05$

or, $x={\cdot}065y-6{\cdot}35$

For $y=600$, we find,

$$x = \cdot065 \times 600 - 6\cdot35$$
$$= 39 - 6\cdot35 = 32\cdot65 \text{ cms.}$$

Example.

For some bivariate data, the following results were obtained :

The mean value of $X = 53\cdot2$,

the mean value of $Y = 27\cdot9$,

the regression coefficient of Y on $X = -1\cdot5$,

and the regression coefficient of X on $Y = -0\cdot2$.

Find (i) the most probable value of Y, when $X = 60$;

(ii) the coefficient of correlation between X and Y.

Regression equation of Y on X is $Y - \bar{Y} = b_{yx}(X - \bar{X})$

or, $Y - 27\cdot9 = -1\cdot5(X - 53\cdot2)$

or, $Y = -1\cdot5X + 1\cdot5 \times 53\cdot2 + 27\cdot9$

or, $Y = -1\cdot5X + 79\cdot80 + 27\cdot9$.

For $X = 60$, we get $Y = -1\cdot5 \times 60 + 79\cdot80 + 27\cdot9$

$$= -90 + 107\cdot7 = 17\cdot7.$$

Again $\gamma = \pm\sqrt{b_{yx} \times b_{xy}} = \pm\sqrt{-1\cdot5 \times -\cdot2} = -\sqrt{\cdot3} = -\cdot547$.

As both the regression coefficients have negative sign, so also γ will have the same sign.

Example.

The following figures relate to years of service and income in hundreds of rupees of the employee of an organisation. Find the initial start for a person applying for a job after having served in another factory for a period of 12 years in a similar capacity.

Length of service (yrs.) :	11	7	9	5	8	6	10
Income (hundreds of Rs.) :	7	5	3	2	6	4	8

Let x represents length of service and y represents income. Here, we are to find initial income after serving 12 yrs. in similar capacity, *i.e.*, to find regression equation of y on x.

Calculation for Regression Equation of y on x

Year x	$x-\bar{x}$ $(\bar{x}=8)$ X	X^2	Income (Rs.) ('00) y	$y-\bar{y}$ $(\bar{y}=5)$ Y	Y^2	XY
11	3	9	7	2	4	6
7	−1	1	5	0	0	0
9	1	1	3	−2	4	−2
5	−3	9	2	−3	9	9
8	3	0	6	1	1	0
6	−2	4	4	−1	1	2
10	2	4	8	3	9	6
$\Sigma x=56$		$\Sigma X^2=28$	$\Sigma y=35$		$\Sigma Y^2=28$	$\Sigma XY=21$

$$\bar{x}=\frac{\Sigma x}{n}=\frac{56}{7}=8\,;\quad \bar{y}=\frac{\Sigma y}{n}=\frac{35}{7}=5$$

Since we are to find regression equation of y on x, we shall use

$$b_{yx}=\frac{\Sigma XY}{\Sigma X^2},$$

$$b_{yx}=\frac{21}{28}=0{\cdot}75.$$

Again to find initial income for the person of 12 years' experience in the same capacity, substituting the respective values we find,

$$y-5=0{\cdot}75(x-8)$$

or, $y-5={\cdot}75x-6$

or, $y={\cdot}75x-1.$

Now, for $x=12$, $y={\cdot}75\times12-1=9-1=8.$

∴ reqd. initial start = Rs. 800.

EXERCISE 10

1. Define 'regression' ; why are there two regression lines ?

2. Distinguish clearly between correlation and regression as concepts used in statistical analysis.

3. The heights (cm.) of a group of fathers and sons are given below :

Ht. of fathers :	158	160	163	165	167	170	167	172	177	181
Ht. of sons :	163	158	167	170	160	180	170	175	172	175

Find the lines of regression and estimate the height of son when the height of the father is 164 cms.

(*Ans.* $y = \cdot69x + 56\cdot6$, $x = \cdot704y + 49\cdot02$, ht. of son $= 166\cdot3$ cms.)

4. Find the regression equation from the following data :

Age of husband :	18	19	20	21	22	23	24	25	26	27
Age of wife :	17	17	18	18	18	19	19	20	21	22

(*Ans.* $x = 1\cdot747y - 10\cdot52$; $y = \cdot527x + 7\cdot04$)

5. From the following table, showing age of cars of a certain make and annual maintenance costs, obtain the regression equation for costs related to age.

Age of cars (yrs.) :	2	4	6	7	8	10	12
Annual maintenance : cost (Rs.)	1600	1500	1800	1900	1700	2100	2000

(*Ans.* $y = 52\cdot86x + 1429\cdot98$)

6. From the following data, obtain the two regression equations :

Sales :	91	97	108	121	67	124	51	73	111	57
Purchases :	71	75	69	97	70	91	39	61	80	47

Hence or otherwise find correlation coefficient between sales and purchases.

[C.A. May '77] (*Ans.* $Y = 0\cdot61X + 15\cdot1$; $X = 1\cdot36Y - 5\cdot2$; $\gamma = +\cdot91$.)

7. From the following data, find the regression equation which you think to be fit :

Age :	56	42	72	36	63	47	55	49	38
Blood pressure :	147	125	160	118	149	128	150	145	115

(*Ans.* age (x) blood pr. (y) (say) $y = 1\cdot27x + 72\cdot86$)

8. The height (inch) and quantity of dry bark (oz.) of 8 sinkano trees are as follows :

Height (x) :	8	11	7	10	12	5	4	6
Quantity (y) :	19	30	25	44	38	25	20	27

Find the regression equation of y on x. If the height is 15 inches, find the quantity of dry bark. (*Ans.* $y = 2\cdot57x + 8\cdot28$; $46\cdot83$ oz.)

9. Find the two regression equations from the following data. If the age of wife is 19 years, find that of husband, and again if the age of husband is 30 years, find that of wife.

Age of husband :	25	22	28	26	35	20	22	40	20	18
Age of wife :	18	15	20	17	22	14	16	21	15	14

(*Ans.* $x = 2{\cdot}23y - 12{\cdot}76$, $y = {\cdot}385x + 7{\cdot}34$, husband's age ($x$), wife's age ($y$) ; 29·6 ; 18·9)

10. A sample of size $n = 16$, yield the following sums : $\Sigma x = 7{\cdot}49$, $\Sigma y = 77{\cdot}90$, $\Sigma y^2 = 454{\cdot}81$, $\Sigma xy = 3156{\cdot}80$ and $\Sigma x^2 = 42{\cdot}177$. Compute the linear regression equation of x on y.
[C. U. B. Com. (Hons.) 1980] (*Ans.* $x = 41{\cdot}22y - 195{\cdot}38$)

11. In a correlation study, the following values are obtained :

	X	Y
Mean	65	67
S. D.	2·5	3·5
Coefficient of correlation	0·8	

Find the two regression equations those are associated with the above values.

(*Ans.* $X = {\cdot}57Y + 26{\cdot}81$; $Y = 1{\cdot}12X - 5{\cdot}8$)

12. You are given the data relating to purchases and sales. Obtain the regression equations by the method of least squares and estimate the likely sales when the purchases equal to 100.

Purchases :	62	72	98	76	81	56	76	92	88	49
Sales :	112	124	131	117	132	96	120	136	97	85

[C. A. Inter, May '75] (*Ans.* $x = 2{\cdot}8 + {\cdot}68y$; $y = 56{\cdot}5 + {\cdot}78x$; 134·5)

13. You are given the following results for the heights (X) and weights (Y) of 1000 business executives :

$\bar{X} = 68$ inches , $\sigma_x = 2{\cdot}5$ inches

$\bar{Y} = 150$ lbs. $\sigma_y = 20$ lbs.

Estimate from the above data :

(i) The weight of a particular executive, who is 5 feet tall.

(ii) The height of a particular executive, whose weight is 200 lbs.

(*Ans.* (i) 111·6 lbs. ; (ii) 71·75 inches)

14. Find σ_y, given that, variance of $X = 36$, $b_{xy} = 0{\cdot}8$, $\gamma = 0{\cdot}5$.

(*Ans.* 3·75)

11

INDEX NUMBERS

Introduction.

Index numbers are devices which indicate by its variations the changes in the magnitude in a group of related variables. The group of related variables may be prices of a specified set of commodities or the volume of production in different sectors or such other concept as intelligence, health or efficiency. They may measure variations over time, over space or between similar categories such as institutions, objects, etc.

The commonest type of such index measures is one known as *Price Index* which measures variation in prices over a span of time. It enables us to know how the price level of a group of commodities has changed at *certain periods of time* as compared with another period, called *base period.* When the price of one commodity rises while the price of another falls and the prices of various commodities all react in different degrees, the index number shall not give here any indication of changes in the values of the individual commodity but will reveal the average net effect of all the changes. Retail Price Index, Wholesale Price Index, Index of Wage-rates (wages being the price of labour) are some of price indices.

Similarly, a quantity index enables us to know the average changes in the quantities of the items belonging to a group of commodities. An Index of industrial production or an Index of the volume of exports are some of the examples of quantity index.

Index numbers enable comparison to be made between the levels of prices and wages, or between the levels of production and wages. Index number also measures the amount of change in the productivity within a firm, or in the value of trade, or the difference in level of intelligence among students of different institutions.

Compilation of Index Numbers.

The most and widely used common type of index is Price Index which measures variation in price level over two different periods of time. As such, the theory will be discussed here in terms of prices over two different periods. The principles involve will of course apply to any other type of index number.

Basic Principle of Construction

(i) Price Relative Method

In constructing the index number, the basic device, in general, is to calculate a set of relatives by expressing the prices of a *given period* as percentage of prices of *base period* and then to average the relatives so derived.

(ii) Aggregate Method

In this method, index number is calculated by expressing the aggregate of the prices of *given period* as a percentage of the aggregate of the prices of *base period.*

Note 1. Price Relative (P.R.) $= \frac{\text{Given period price}}{\text{Base period price}} \times 100.$

Note 2. Simple Index Number is an index number which measures change in a single item.

Problems in the Construction of Index Numbers.

We are generally confronted with the following five major problems while constructing any Index Number :

(i) Definition of Purpose of the Index Number.

(ii) Selection of Items and Collection of Data.

(iii) Selection of Base Period.

(iv) Selection of Weights.

(v) Choice of Average.

(i) *Definition of Purpose of the Index Number*

The first essential point to be considered is a precise statement of the purpose for which a particular index number is to be constructed. All index numbers will not serve the same purpose or there is no all-purpose index number. The purpose of construction will help the selection of items, the selection of base period, the selection of weights, etc. Thus, for compilation of consumers price index number wholesale prices should not be included. To study the changes in general price level after independence the base year will be year immediately preceding the year of independence. For construction of index number of building materials item like cement is more important than the item glass and hence the item cement will receive high weightage than that of glass.

As the purpose will determine the data to be collected, the data available or lack of it, may necessitate the modification of purpose.

(ii) *Selection of Items and Collection of Data*

Items selected for the construction of index number should be relevant. Haphazard selection will not only be confusing but sometimes useless. In constructing consumers price index for working-class items like T. V., Refrigerator, Motor-car should not be taken into consideration.

Index number takes into account a large number of items. But it might not be practicable to include all the relevant items in view of cost and time. Hence, there is the necessity of sampling. Items selected should be representative of all the relevant items.

Items included should neither be too large nor too few. If the items included be too large, the more will the index number be representative but it will create difficulty in the accurate collection of data, apart from the increased cost and time. Again, inclusion of too few items in the index number will make it unreliable.

Hence, items selected should be both *adequate in number* and *representative* in character.

Sometimes, the prices of one and the same item becomes incomparable between different dates due to considerable change in its quality. Hence, items of standard quality should be entered in the index number construction.

Again, for same item the price quotations may vary from place to place and for quality to quality and since all prices at all transactions and of all varieties cannot be taken into account, *samples of prices of all varieties and at all transactions should be taken in such a manner that they will represent adequately the overall situation.*

As the tastes and habits change with the passage of time, some new items become important while some old items become obsolete and hence the selection of items for preparing the index number is to be very carefully done making necessary additions and alterations.

The data collected should be basically *accurate, unbiased* and *dependable* and may be collected from published sources or by appointing some persons or institutions who can supply them as and when required. The accuracy of the data is to be checked properly before use.

(iii) *Selection of Base Period*

The base period must be a period of *economic stability.* The prices in this period must not be fixed by law, this period must be one in which no abnormal increase or decrease is found.

The problem of choice of a base period having economic stability is very difficult. To solve this, aggregate or average of some periods may be taken as base so that abnormalities in one direction will neutralize the abnormalities in another direction.

Base period should neither be too long, nor too short. Generally, a year is taken as base. A week, or a month, or a group of years or months or weeks may also be taken as base.

Base period should not be too distant from the given period as this may cause a change in the importance of some items. Further, the quality of the items may differ substantially if the interval is too wide and the prices at the two periods also become incomparable. The more and more the base period is away from the given period, the relative change will have larger and larger variability and hence shifting of the base period to some other period which is not too distant from the given period becomes necessary.

(iv) *Selection of Weights*

In constructing the index numbers all the items included are not of equal importance in the sense that a similar change in the prices of *different items* does not affect the price level to the *same extent*. Therefore, it is necessary that due consideration be given to the relative importance of different items. This is done by allocating weights. Thus, we have two types of indices : (1) Simple or Unweighted index and (2) Weighted index. In simple aggregate index, all the items are considered equally important, hence weight assigned to each item is the same. Similarly in simple average of relatives the weight assigned to each relative is the same. So, in reality the simple index numbers are *arbitrarily* weighted index numbers. But as it is reasonable that each item included should be allowed to have due importance on the index number, it should be *deliberately* weighted. To construct a weighted index number, generally, price relatives are weighted by values, prices by quantities and quantities by prices. The prices and quantities used as weights may be the prices or quantities of (1) base period, (2) given period, (3) average or total of base and given periods, (4) average of all the periods included in the index number or (5) average of several periods thought to be typical.

(v) *Choice of Average*

The choice of average will normally be dictated by practical consideration. In theory it is possible to use any form of average. Arithmetic mean is used in most of the important indices due to its intelligibility and simplicity in calculation. Some, however, use Geometric mean for it is mathematically suitable and is less susceptible to a few items of very high and low values. Median and Mode are generally erratic and other averages are complicated.

Methods of Constructing Index Numbers.

The following are the different methods of construction of Index Numbers :

(1) METHOD OF AGGREGATES

(a) *Simple aggregate of Prices.*

(b) *Weighted aggregate of Prices.*

(2) METHOD OF RELATIVES

(a) *Simple average of Price Relatives.*

(b) *Weighted average of Price Relatives.*

Symbols and Notations to be used throughout this chapter :

$p_o', p_o'', p_o''', \ldots\ldots\ldots$ represent prices in the base period denoted by the suffix o.

$q_o', q_o'', q_o''', \ldots\ldots\ldots$ represent quantities in the base period denoted by the suffix o.

$p_n', p_n'', p_n''', \ldots\ldots\ldots$ represent prices in the given period denoted by suffix n.

$q_n', q_n'', q_n''', \ldots\ldots\ldots$ represent quantities in the given period denoted by the suffix n.

$p_o' + p_o'' + p_o''' + \cdots\cdots = \Sigma p_o =$ Total of base period prices.

$q_o' + q_o'' + q_o''' + \cdots\cdots = \Sigma q_o =$ Total of base period quantities.

Similarly $\Sigma p_n = p_n' + p_n'' + p_n''' + \cdots\cdots$

$\Sigma q_n = q_n' + q_n'' + q_n''' + \cdots\cdots$

$\Sigma p_o q_o = p_o' q_o' + p_o'' q_o'' + p_o''' q_o''' + \cdots\cdots$

$\Sigma p_n q_n = p_n' q_n' + p_n'' q_n'' + p_n''' q_n''' + \cdots\cdots$

$I_{o,n} =$ Index Number with base period o and given period n.

We shall, however, use I in place of $I_{o,n}$ when there is no possibility of confusion.

(1-a) *Simple Aggregate of Prices*

In this method, the Price Index will be the aggregate of prices of the given period expressed as a percentage of that in the base period.

Symbolically, $I = \dfrac{\Sigma p_n}{\Sigma p_o} \times 100.$ ... ... (1)

EXAMPLE :

From the data given below, calculate the index number of prices for all the years with reference to 1970 as the base year using Simple Aggregate Method.

Items	*Prices*			
	1970	1971	1972	1973
Rice	56·4	58·7	57·2	60·4
Wheat	48·2	50·3	51·7	53·3
Pulse	121·3	124·6	130·3	133·5

SOLUTION :

$$\text{Simple Aggregate Index Number} = \frac{\Sigma p_n}{\Sigma p_0} \times 100.$$

Let p_0, p_1, p_2, p_3 represent prices in 1970, 1971, 1972 and 1973 respectively.

Items	p_0	p_1	p_2	p_3
Rice	56·4	58·7	57·2	60·4
Wheat	48·2	50·3	51·7	53·3
Pulse	121·3	124·6	130·3	133·5
Total	225·9	233·6	239·2	247·2

$$\therefore \quad \Sigma p_0 = 225{\cdot}9, \ \Sigma p_1 = 233{\cdot}6, \ \Sigma p_2 = 239{\cdot}2, \ \Sigma p_3 = 247{\cdot}2$$

Price Index Number for the year 1971 with 1970 as base year

$$I_{0,1} = \frac{\Sigma p_1}{\Sigma p_0} \times 100 = \frac{233{\cdot}6}{225{\cdot}9} \times 100 = 103{\cdot}4$$

Price Index Number for 1972 with 1970 as base year

$$I_{0,2} = \frac{\Sigma p_2}{\Sigma p_0} \times 100 = \frac{239{\cdot}2}{225{\cdot}9} \times 100 = 105{\cdot}9$$

Similarly, $I_{0,3} = \dfrac{\Sigma p_3}{\Sigma p_0} \times 100 = \dfrac{247{\cdot}2}{225{\cdot}9} \times 100 = 109{\cdot}4.$

EXAMPLE :

From the following data, construct an index for 1978 taking 1970 as base using Simple Aggregate Method

	1970	1978
Price of Rice per kg.	Rs. 2·20	Rs. 3·45
Price of Egg per doz.	Rs. 3·10	Rs. 4·50.

SOLUTION :

Construction of Price Index :

	1970	1978
Price of Rice per kg.	Rs. 2·20	Rs. 3·45
Price of Egg per dozen	Rs. 3·10	Rs. 4·50
Total	Rs. 5·30	Rs. 7·95

$$I_{o,n}=\frac{7{\cdot}95}{5{\cdot}30}\times100=150.$$

Now, instead of taking price of rice per kg., if we take the price of rice per quintal, we have,

	1970	1978
Price of Rice per quintal	Rs. 220·00	Rs. 345·00
Price of Egg per dozen	Rs. 3·10	Rs. 4·50
Total	Rs. 223·10	Rs. 349·50

$$\therefore\ I_{o,n}=\frac{349{\cdot}50}{223{\cdot}10}\times100=156{\cdot}2.$$

It is now seen that index number has been increased from 150 to 156·2 as we take the price of rice per quintal instead of price of rice per kg. Thus, this method has a serious defect as a difference in the unit of quotation makes a difference in the index. So, manipulation becomes possible by quoting prices in different units and which is not desirable. As a matter of fact, equal importance is given to all the items here which is not correct. For these reasons the simple aggregative index is rarely used.

(1-*b*) *Weighted Aggregate of Prices*

This price index is the aggregate of weighted prices of the given period expressed as a percentage of that of the base period.

Symbolically, $I_{o,n}=\dfrac{\Sigma p_n w}{\Sigma p_o w}\times100,$ ... ... (2)

Where $w', w'', w''', \ldots$ are the weights of different items.

$$\Sigma p_n w=p_n'w'+p_n''w''+p_n'''w'''+\cdots$$

and $\Sigma p_o w=p_o'w'+p_o''w''+p_o'''w'''+\cdots$

In general, the weights used in this method are *quantities consumed, marketed or sold* in the base period or given period or the average of several periods.

Various formulas obtained by using various system of weights :

(A) Laspeyres' Formula

In this formula base period quantities are used as weights. So, substituting $w = q_0$ in (2) we have,

$$I_{0,n} = \frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \times 100. \quad \ldots \quad (3)$$

(B) Paasche's Formula

Here given period quantities are used as weights. So, putting q_n for w in (2) we obtain

$$I_{0,n} = \frac{\Sigma p_n q_n}{\Sigma p_0 q_n} \times 100. \quad \ldots \quad \ldots \quad (4)$$

From the practical point of view the Laspeyres' Method is easier to use since in this method, base year quantities being weights, it is not necessary to select a new sets of weights as the given period changes. On the other hand, Paasche's Method involves the selection of new quantity weights for each period considered and in most cases the weights are difficult to obtain so frequently.

It is most common that, if the demand schedules of the consumers are fixed, the consumers purchase relatively larger quantities of those articles that have decreased in price relatively to other articles and relatively smaller quantities of those articles that have increased in price relatively to other articles. For this reason it is possible that Laspeyres' Index which uses base period quantites as weights will show an increase and Paasche's Index which uses given period quantities as weights will show a decrease in the price level. Thus Laspeyres' Index is said to have an *upward bias* and represents the *upper limit* of the price change while Paasche's Index has a *downward bias* and represent *lower limit* of the price change.

(C) Marshall-Edgeworth Formula

Since Laspeyres' formula has an upward bias and Paasche's formula has a downward bias, the compromise solution known as Marshall-Edgeworth formula which uses the average of the base year quantities and given year quantities as weights and is

$$I_{0,n} = \frac{\Sigma p_n(q_0 + q_n)/2}{\Sigma p_0(q_0 + q_n)/2} \times 100 = \frac{\Sigma p_n q_0 + \Sigma p_n q_n}{\Sigma p_0 q_0 + \Sigma p_0 q_n} \times 100. \quad \ldots \quad (5)$$

(D) Fisher's Ideal Formula

Another compromise solution called the Fisher's Ideal Formula which is the Geometric Average of Laspeyres' formula and Paasche's formula and is given by

$$I_{0,n} = \sqrt{\frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_n q_n}{\Sigma p_0 q_n}} \times 100. \quad \ldots \quad \ldots \quad (6)$$

This formula is known as *Ideal*, since the upward bias of Laspeyres' index number neutralizes to a great extent by the downward bias of Paasche's index number.

EXAMPLE :

Construct the index number for 1960 using 1959 as base year from the following data using Weighted Aggregate Method :

Items	*Weights*	*Price per unit* (*Rs.*)	
		1959	1960
A	40	16·00	20·00
B	25	40·00	60·00
C	5	0·50	0·50
D	20	5·12	6·25
E	10	2·00	1·50

SOLUTION :

$$\text{Weighted Aggregate Index Number} = \frac{\Sigma wp_n}{\Sigma wp_0} \times 100.$$

Let p_0 and p_n represent prices in 1959 and 1960 respectively and w represents weights.

Table for Calculation of Index Number

Items	w	p_0	p_n	wp_0	wp_n
A	40	16·00	20·00	640·00	800·00
B	25	40·00	60·00	1000·00	1500·00
C	5	0·50	0·50	2·50	2·50
D	20	5·12	6·25	102·40	125·00
E	10	2·00	1·50	20·00	15·00

$$\therefore \quad \Sigma wp_0 = 1764{\cdot}90, \ \Sigma wp_n = 2442{\cdot}50$$

$$\text{Index Number} = \frac{2442{\cdot}50}{1764{\cdot}90} \times 100 = 138{\cdot}4.$$

EXAMPLE :

Construct Laspeyres', Paasche's, Edgeworth-Marshall's and Fisher's Ideal Index Numbers from the data shown in the p. 298.

Items	*Quantity*		*Price*	
	Base year	*Current year*	*Base year*	*Current year*
A	10	12	12	15
B	5	10	8	10
C	12	16	10	12

SOLUTION :

Table for Calculation of Index Numbers

Items	q_0	q_n	p_0	p_n	p_0q_0	p_nq_0	p_0q_n	p_nq_n
A	10	12	12	15	120	150	144	180
B	5	10	8	10	40	50	80	100
C	12	16	10	12	120	144	160	192

$\therefore \quad \Sigma p_0q_0 = 280, \ \Sigma p_nq_0 = 344, \ \Sigma p_0q_n = 384, \ \Sigma p_nq_n = 472$

Laspeyres' Index Number $= \frac{344}{280} \times 100 = 122{\cdot}86.$

Paasche's Index Number $= \frac{472}{384} \times 100 = 122{\cdot}92.$

$$\text{Edgeworth-Marshall's Index Number} = \frac{344+472}{280+384} \times 100,$$

$$= \frac{816}{664} \times 100 = 122{\cdot}9.$$

Fisher's Ideal Index Number $= \sqrt{122{\cdot}86 \times 122{\cdot}92} = 122{\cdot}88.$

EXAMPLE :

Calculate the Laspeyres', Paasche's and Edgeworth-Marshall's Index Numburs from the following data :

Items	*Base year price* (p_0) (Rs.)	*Base year quantity* (q_0) (Kg.)	*Current year price* (p_n) (Rs.)	*Current year quantity* (q_n) (Kg.)
A	5	50	10	56
B	3	100	4	120
C	4	60	6	60
D	11	30	14	24
E	7	40	10	36

SOLUTION :

Calculation of Index Numbers

Items	p_0	q_0	p_n	q_n	p_0q_0	p_0q_n	p_nq_0	p_nq_n
A	5	50	10	56	250	280	500	560
B	3	100	4	120	300	360	400	480
C	4	60	6	60	240	240	360	360
D	11	30	14	24	330	264	420	336
E	7	40	10	36	280	252	400	360

$\therefore \quad \Sigma p_0q_0 = 1400, \ \Sigma p_0q_n = 1396, \ \Sigma p_nq_0 = 2080, \ \Sigma p_nq_n = 2096$

$$\text{Laspeyres' Index Number} = \frac{2080}{1400} \times 100 = 148{\cdot}6$$

$$\text{Paasche's Index Number} = \frac{2096}{1396} \times 100 = 150{\cdot}1$$

$$\text{Edgeworth-Marshall's Index Number} = \frac{2080+2096}{1400+1396} = \frac{4176}{2796} \times 100 = 149.$$

(2-*a*) *Simple Average of Price Relatives*

It is the average of a set of price relatives obtained by expressing the given period prices as percentage of the corresponding prices of the base period.

(i) *When A. M. is used :*

$$I_{0,n} = \frac{\sum \frac{p_n}{p_0}}{N} \qquad \cdots \qquad \cdots \quad (7)$$

(ii) *When G. M. is used :*

$$I_{0,n} = \left(\frac{p_n'}{p_0'} \times \frac{p_n''}{p_0''} \times \frac{p_n'''}{p_0'''} \times \cdots\cdots\right)^{\frac{1}{N}} \qquad \cdots \qquad \cdots \quad (8)$$

where N is the number of commodities.

(2-*b*) *Weighted Average of Price Relatives*

If $w', w'', w''', \cdots\cdots$ be the weights, then the weighted average of price relatives are obtained by multiplying each relatives by its weights and dividing their sum by the sum of the weights. The formula is as follows :

$$I_{0,n}=\frac{\sum\left(\frac{p_n}{p_o}\times w\right)}{\Sigma w}\times 100. \qquad \cdots \qquad \cdots \quad (9)$$

Here, in general, the weights are value weights.

(i) *When the weights are base period values*, then $w=p_oq_o$ and we obtain from (9),

$$I_{0,n}=\frac{\sum\left(p_oq_o\times\frac{p_n}{p_o}\right)}{\Sigma p_oq_o}\times 100 \qquad \cdots \qquad \cdots \quad (10)$$

$$=\frac{\Sigma p_nq_o}{\Sigma p_oq_o}\times 100$$

which is nothing but Laspeyres' Formula.

(ii) *When weights are given period quantities at base prices*, then $w=p_oq_n$ and from (9) we get,

$$I_{0,n}=\frac{\sum\left(\frac{p_n}{p_o}\times p_oq_n\right)}{\Sigma p_oq_n}\times 100 \qquad \cdots \qquad \cdots \quad (11)$$

$$=\frac{\Sigma p_nq_n}{\Sigma p_oq_n}\times 100$$

which is Paasche's Formula.

Construction of General Index Number from Group Indices.

When the index numbers of various groups of items are given, then usually the weighted arithmetic mean of the separate index numbers will give the index number of all the groups combined. So, if $I_1, I_2, I_3, \ldots\ldots, I_n$ be the index numbers of various groups and $w_1, w_2, w_3, \ldots\ldots, w_n$ be the corresponding weights, then the combined index number of all the groups is given by

(i) *When A. M. is used :*

$$I=\frac{I_1w_1+I_2w_2+\cdots\cdots+I_nw_n}{w_1+w_2+\cdots\cdots+w_n}.$$

(ii) *When G. M. is used :*

$$I=(I_1^{w_1}\,.\,I_2^{w_2}\cdots\cdots I_n^{w_n})^{\frac{1}{w_1+w_2+\cdots+w_n}}$$

where n = number of various groups.

EXAMPLE :

Given the following data compute price index number for the year 1978 (base year 1970) using (i) simple average, (ii) weighted average of price relatives.

Items	*Price*		*Weights*
	1970	1978	
Rice	180	200	11
Wheat	140	170	5
Pulse	480	500	4
Fish	14	16	1

SOLUTION :

Calculation of Price Relatives

Items	p_o	p_n	w	$\frac{p_n}{p_o}$	$\frac{p_n}{p_o}w$
Rice	180	200	11	111·1	1222·1
Wheat	140	170	5	121·4	607·0
Pulse	480	500	4	104·2	416·8
Fish	14	16	1	114·2	114·2

$$\therefore \quad \Sigma\frac{p_n}{p_o}=450{\cdot}9,\ \Sigma\frac{p_n}{p_o}w=2360{\cdot}1,\ \Sigma w=21.$$

Simple average of Price Relative Index Number $=\frac{450{\cdot}9}{4}=112{\cdot}7.$

Weighted average of Price Relative Index Number $=\frac{2360{\cdot}1}{21}=112{\cdot}4.$

EXAMPLE :

From the following compute price index number using simple average of price relatives based on (i) A.M. and (ii) G.M :

Commodities	*Price*	
	Base year	*Current year*
A	25	30
B	20	22
C	30	33
D	12	15
E	90	99

SOLUTION :

Calculation of Price Index Numbers

Commodities	p_0	p_n	$\frac{p_n}{p_0}$	log p_0	log p_n	log p_n − log p_0
A	25	30	30÷25	1·3979	1·4771	0·0792
B	20	22	22÷20	1·3010	1·3424	0·0414
C	30	33	33÷30	1·4771	1·5185	0·0414
D	12	15	15÷12	1·0792	1·1761	0·0969
E	90	99	99÷90	1·9542	1·9956	0·0414
Total						0·3003

Simple average price relative
Index Number (using A.M.) $= \frac{1}{5}\left(\frac{30}{25}+\frac{22}{20}+\frac{33}{30}+\frac{15}{12}+\frac{99}{90}\right)\times 100 = 113.$

Simple average price relative
Index Number (using G.M.) $= \text{I} = \left\{\frac{p_n{}^{\text{I}}}{p_0{}^{\text{I}}}\times\frac{p_n{}^{\text{II}}}{p_0{}^{\text{II}}}\times\frac{p_n{}^{\text{III}}}{p_0{}^{\text{III}}}\times\frac{p_n{}^{\text{IV}}}{p_0{}^{\text{IV}}}\times\frac{p_n{}^{\text{V}}}{p_0{}^{\text{V}}}\right\}^{\frac{1}{5}}$
$\times 100.$

$$\therefore \quad \text{Log I} = \frac{1}{5}\,\Sigma \log\frac{p_n}{p_0} + \log 100$$
$$= \tfrac{1}{5}\,\Sigma(\log p_n - \log p_0) + \log 100$$
$$= \tfrac{1}{5}\times 0{\cdot}3003 + 2$$
$$= 0{\cdot}06006 + 2$$
$$= 2{\cdot}06006$$
$$\text{I} = 114{\cdot}8.$$

Chain-base Method of Construction of Index Numbers.

In this method indices are compiled for each year with previous year as base and chain them, so as to get a series referring back to a base year. Symbolically,

$$I_{0,n}' = I_{0,1} \times I_{1,2} \times \cdots \times I_{(n-1),\,n}$$

where $I_{i,j}$ the index with i as base and j as given year. $I_{0,1}$, $I_{1,2}$, $I_{2,3}$, $\cdots$ are called *Link Indices* and $I_{0,n}'$ is called *Chain Index.*

If Laspeyres' formula is used then, *omitting the factor 100*, we have,

$$I_{0,1} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0}\,;\; I_{1,2} = \frac{\Sigma p_2 q_1}{\Sigma p_1 q_1}\,;\; \cdots\; I_{(n-1),n} = \frac{\Sigma p_n q_{n-1}}{\Sigma p_{n-1} q_{n-1}}$$

$$\text{and} \quad I_{0,n}' = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_2 q_1}{\Sigma p_1 q_1} \times \cdots \times \frac{\Sigma p_n q_{n-1}}{\Sigma p_{n-1} q_{n-1}}.$$

And when Paasche's form is used then, *omitting the factor 100*, obviously,

$$I_{0,1} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}\,;\; I_{1,2} = \frac{\Sigma p_2 q_2}{\Sigma p_1 q_2}\,;\; \cdots\; I_{(n-1),n} = \frac{\Sigma p_n q_n}{\Sigma p_{n-1} q_n},$$

$$\text{and} \quad I_{0,'n} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma p_2 q_2}{\Sigma p_1 q_2} \times \cdots \times \frac{\Sigma p_n q_n}{\Sigma p_{n-1} q_n}.$$

Advantages :

(i) Direct comparison between two successive periods possible through link indices which is more important to commercial users than the comparison through remote base period.

(ii) When weights are changing rapidly it is desirable to use chain-base indices.

(iii) It facilitates introduction of new item replacing the obsolete old one.

Disadvantages :

(i) Amount of calculation involved in this method is immense.

(ii) Easy interpretation is lacking.

EXAMPLE :

From the following prices (Rs.) of the groups of items, calculate chain indices with 1974 as the base year, using A.M. of price relatives ;

Group	1974	1975	1976	1977	1978
1	3	4	5	6	7
2	8	10	12	15	18
3	6	8	5	10	12

SOLUTION :

Calculation of price relatives based on preceding year

Group	1974	1975	1976	1977	1978
1	100	133·33	125·00	120·00	116·67
2	100	125·00	120·00	125·00	120·00
3	100	133·33	62·50	200·00	120·00
Total :	300	391·66	307·50	445·00	356·67
Link Index :	100	130·55	102·50	148·33	118·89

Calculation of Chain Indices

Year	Link Index	Chain Index
1974	100	100
1975	130·55	$\frac{100}{100} \times 130{\cdot}55 = 130{\cdot}55$
1976	102·50	$\frac{130{\cdot}55}{100} \times 102{\cdot}50 = 133{\cdot}81$
1977	148·33	$\frac{133{\cdot}81}{100} \times 148{\cdot}33 = 198{\cdot}78$
1978	118·89	$\frac{198{\cdot}78}{100} \times 118{\cdot}89 = 237{\cdot}42$

Note. Fixed-base Method and Chain-base Method of Index Numbers when Paasche's formula is used.

Fixed-base Method

$$I_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}$$

$$I_{02} = \frac{\Sigma p_2 q_2}{\Sigma p_0 q_2}$$

....

$$I_{0n} = \frac{\Sigma p_n q_n}{\Sigma p_0 q_n}$$

Chain-base Method

$$I_{01}' = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}$$

$$I_{02}' = I_{01} \times I_{12} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma p_2 q_2}{\Sigma p_1 q_2}$$

....

$$I_{0n}' = I_{01} \times I_{12} \times \cdots \times I_{(n-1)n}$$

$$= \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma p_2 q_2}{\Sigma p_1 q_2} \times \cdots \times \frac{\Sigma p_n q_n}{\Sigma p_{n-1} q_n}.$$

Consumer Price Index or Cost of Living Index.

Cost of Living Index Numbers are generally intended to represent the average change in prices over a period of time, paid by the ultimate consumers for a fixed set of goods and services. It measures the relative changes over time in the cost level required to maintain similar standard of living of a specified class of people, whose consumption pattern does not vary widely, living in a specified region within which retail prices are almost equal.

Items contributing to Consumer Price Index are generally classified under five major groups :

(i) *Food*, (ii) *Clothing*, (iii) *Fuel and Light*, (iv) *Housing*, (v) *Miscellaneous*.

Each of the groups again includes a large number of items. For example, the group *Food* includes Rice, Wheat, Milk, Vegetables, Egg, Fish/Meat, Tea, Butter/Ghee, etc. The items of consumption included are those which the people for whom the index is meant generally consume. However, for saving time and labour, the number of items selected should be limited but representative. The retail price quotations of the items selected must be easily available and should be taken at regular intervals from those sources from which the people obtain their goods and services. For each item if the price-quotations obtained from different sources differ then the average of the price-quotations to be used.

Relative importance of various items for different classes of people being different, cost of living index should always be weighted. Index Numbers are then calculated for each of the five groups by any of the following methods :

(i) *Aggregate Expenditure Method or Aggregative Method.*

(ii) *Family Budget Method or Method of Weighted Relatives.*

(i) *Aggregate Expenditure Method* : In this method quantities of items cousumed by particular group in the base year or their proportions constitute the weights and the Index Number is the aggregate of weighted prices of given year expressed as a percentage of that in the base year.

Symbolically, Group Index $(I)=\frac{\Sigma p_n q_0}{\Sigma p_0 q_0}\times 100$

which is the Laspeyres' Method discussed earlier.

(ii) *Family Budget Method* : In this method the expenditure of an average family on various items is to be estimated after carefully studying the *Family Budgets* of large number of people for whom the Index is to be compiled. This 'family budget inquiry' is conducted in the *base year*, using the technique of random sampling. The expenditure of an average family on each item constitutes the weight (w) of the item within the group. The weights are thus the value weights (i.e., p_0q_0) and the Index Number is calculated by using weighted average of price relatives. Thus,

$$\text{Groupe Index (I)} = \frac{\sum\left(\frac{p_n}{p_0} \times 100\right)w}{\Sigma w}$$

$$= \frac{\sum\left(\frac{p_n}{p_0} \times 100\right)p_0q_0}{\Sigma p_0q_0}, \qquad \text{since } w = p_0q_0.$$

$$= \frac{\Sigma p_nq_0}{\Sigma p_0q_0} \times 100.$$

This method is thus the same as that of the Aggregative Method.

In practice, however, the expenditure of an average family on each item is expressed as a percentage of the total expenditure of the group and the percentage so derived is used as weight (w) of the item within the group. Hence, in this case,

$$w = \frac{p_0q_0}{\Sigma p_0q_0} \times 100, \text{ and we get,}$$

$$\text{Group Index (I)} = \frac{\sum w\left(\frac{p_n}{p_0} \times 100\right)}{100}.$$

The Cost of Living Index Number is then calculated using Weighted Arithmetic Average of Group Indices,

$$\text{C. L. I.} = \frac{\Sigma \text{IW}}{\Sigma \text{W}}$$

where W, the weight of a group index, is the percentage of the total expenditure of an average family spent on that group.

Uses of Cost of Living Index Numbers.

(1) C. L. I. numbers are used for adjustment of dearness allowance to maintain the same standard of living as in the base date.

(2) It is used in fixing the wage policy, taxation and in a large number of economic policy.

(3) Reciprocal of C. L. I. is used as a reliable measure of the purchasing power of money. In fact, the purchasing power of money varies inversely with the C. L. I.

(4) Real wages can be measured by dividing the actual wage received during a period by the corresponding C. L. I. of that period.

EXAMPLE :

The following table gives the group index numbers and the corresponding group weights with reference to the cost of living for a given year. Construct overall cost of living index for the year.

Groups	*Index Number*	*Weight*
Food	360	60
Clothing	295	5
Fuel and Light	287	7
House rent	110	8
Miscellaneous	315	20

SOLUTION :

Calculation for C. L. I.

Groups	*Index Number (I)*	*Weight (W)*	IW
Food	360	60	21600
Clothing	295	5	1475
Fuel and Light	287	7	2009
House rent	110	8	880
Miscellaneous	315	20	6300
Total		100	32264

$$\therefore \quad \text{Cost of Living Index Number} = \frac{32264}{100} = 322{\cdot}64.$$

EXAMPLE :

Calculate the changes in the Cost of Living figures for 1978 as compared with 1975.

Items	*Food*	*Rent*	*Clothing*	*Fuel*	*Miscellaneous*
Prices (1975) :	250	60	80	50	200
Prices (1978) :	270	80	100	50	250
Percentage expenditure :	35	20	15	10	20

It is decided by the management of a firm to increase the D.A. of the workers, who were drawing wages Rs. 200 per month per worker, in 1975. How much D.A. should be given to them, so that they are compensated on account of change in C. L. I. for the year 1978.

SOLUTION :

Let p_0 and p_n represent the prices for the years 1975 and 1978 and W represents the percentage expenditure.

Construction of C. L. I. for 1978 with 1975 as base year

Items	W	p_0	p_n	$P=\frac{p_n}{p_0}\times 100$	*PW*
Food	35	250	270	108·00	3780·00
Rent	20	60	80	133·33	2666·60
Clothing	15	80	100	125·00	1875·00
Fuel	10	50	50	100·00	1000·00
Miscellaneous	20	200	250	125·00	2500·00
Total	100				11821·60

$$\text{Cost of Living Index Number} = \frac{11821·60}{100} = 118·22.$$

Thus there is an increase of 18·22% in the cost of living in 1978 as compared to 1975.

Calculation of D.A.

Since the C. L. I. is 118·22, the worker, who was drawing Rs. 100 in 1975, at present he should draw Rs. 118·22. As the worker was drawing Rs. 200 in 1975, at present he should draw

$$\frac{118·22}{100}\times 200 = \text{Rs. } 236·44.$$

Hence the D.A. should be increased by Rs. 236·44 – Rs. 200·00 = Rs. 36·44 to compensate on account of change in C. L. I.

Quantity Index Numbers.

A Quantity Index Number measures the change in volume of quantity produced, consumed or distributed. Problems and methods of construction of this index number are similar to those involved in price index number. So this will not be discussed separately. The quantity index number formulae can be obtained easily by changing p to q and q to p in the corresponding price index number formulae.

Some important Quantity Index Number formulae

Price Index Numbers	*Corresponding Quantity Index Numbers*
(1) $P_{o,n}=\frac{\Sigma p_n}{\Sigma p_o}\times 100$	(1) $Q_{o,n}=\frac{\Sigma q_n}{\Sigma q_o}\times 100$
(2) $P_{o,n}=\frac{\sum\frac{p_n}{p_o}}{n}\times 100$	(2) $Q_{o,n}=\frac{\sum\frac{q_n}{q_o}}{n}\times 100$
(3) $P_{o,n}=\frac{\Sigma p_nq_o}{\Sigma p_oq_o}\times 100$	(3) $Q_{o,n}=\frac{\Sigma q_np_o}{\Sigma q_op_o}\times 100$
(4) $P_{o,n}=\frac{\Sigma p_nq_n}{\Sigma p_oq_n}\times 100$	(4) $Q_{o,n}=\frac{\Sigma q_np_n}{\Sigma q_op_n}\times 100$
(5) $P_{o,n}=\sqrt{\frac{\Sigma p_nq_o}{\Sigma p_oq_o}\times\frac{\Sigma p_nq_n}{\Sigma p_oq_n}}\times 100$	(5) $Q_{o,n}=\sqrt{\frac{\Sigma q_np_o}{\Sigma p_oq_o}\times\frac{\Sigma p_nq_n}{\Sigma p_nq_o}}\times 100$

EXAMPLE :

Given the following data, calculate quantity index numbers using (i) Laspeyres' formula, (ii) Paasche's formula and (iii) Fisher's formula.

Items	*Base year price* (p_o) (*Rs.*)	*Base year quantity* (q_o) (*Kg.*)	*Current year price* (p_n) (*Rs.*)	*Current year quantity* (q_n) (*Kg.*)
A	5	50	10	56
B	3	100	4	120
C	4	60	6	60
D	11	30	14	24
E	7	40	10	36

SOLUTION :

Calculations for Index Numbers

Items	p_o	q_o	p_n	q_n	p_oq_o	p_nq_n	p_nq_o	p_oq_n
A	5	50	10	56	250	560	500	280
B	3	100	4	120	300	480	400	360
C	4	60	6	60	240	360	360	240
D	11	30	14	24	330	336	420	264
E	7	40	10	36	280	360	400	252

$\Sigma p_oq_o = 1400$, $\Sigma p_oq_n = 1396$, $\Sigma p_nq_o = 2080$, $\Sigma p_nq_n = 2096$.

$$\text{Laspeyres' Quantity Index Number} = \frac{1396}{1400} \times 100 = 99{\cdot}7.$$

$$\text{Paasche's Quantity Index Number} = \frac{2096}{2080} \times 100 = 100{\cdot}8.$$

$$\text{Fisher's Ideal Quantity Index Number} = \sqrt{99{\cdot}7 \times 100{\cdot}8} = 100{\cdot}2.$$

EXAMPLE :

Calculate the Production Index for 1921 with 1910 as base year.

Kind of fuel	*Quantities*		*Value in* 1921
	1910	1921	(million rupees)
Bituminous Coal (million tons)	417·10	415·90	1948
Anthracite Coal (million tons)	84·49	90·44	731
Oil (million barrels)	209·60	472·20	712

SOLUTION :

Let q_o be the base year quantity and q_n be the current year quantity. Then value in 1921, *i.e.*, current year is p_nq_n. Since p_nq_n is given, Paasche's quantity index number formula may be used. For this we are also to find the product p_nq_o. Now, we can find p_n using the relation $p = \frac{p_nq_n}{q_n} = \frac{\text{Value in 1921}}{\text{Quantity in 1921}}$.

Calculation of Production Index Number

Items	q_o	q_n	p_nq_n	$p_n = \frac{p_nq_n}{q_n}$	p_nq_o
Bituminous coal (million tons)	417·10	415·90	1948	4.68	1952·03
Anthracite coal (million tons)	84·49	90·44	731	8·08	682·68
Oil (million barrels)	209·60	472·20	712	1·52	318·59

$\Sigma p_nq_n = 3391·00$, $\Sigma p_nq_o = 2953·30$.

Paasche's Quantity Index Number $= \frac{3391·00}{2953·30} \times 100 = 114\ 82$

Tests of Index Numbers.

To judge the merit of various index number formulae noted economists and statisticians have suggested a number of mathematical tests. Of these Time Reversal Test, Factor Reversal Test and Circular test are most important. An index may be considered 'ideal' if it meets these tests.

(1) *Time Reversal Test.*

"The test is that the formula for calculating the index number should be such that it will give the same ratio between one point of comparison and the other, no matter which of the two is taken as base." Symbolically, $I_{o,n} \times I_{n\,o} = 1$, *i.e.*, $I_{o,n} = \frac{1}{I_{n\cdot o}}$, *i.e.*, $I_{o,n}$ is the reciprocal of $I_{n\,o}$, where $I_{o,n}$ is the index for time '*n*' on time '*o*' as base and $I_{n\,o}$ is the index for time '*o*' on time '*n*' as base.

Unweighted G.M. of price relatives satisfies this test for :

$$I_{o,n} = \left(\frac{p_n'}{p_o'} \times \frac{p_n''}{p_o''} \times \frac{p_n'''}{p_o'''} \times \cdots\right)^{\frac{1}{N}}$$

is the reciprocal of

$$I_{n\,o} = \left(\frac{p_o'}{p_n'} \times \frac{p_o''}{p_n''} \times \frac{p_o'''}{p_n'''} \times \cdots\right)^{\frac{1}{N}}$$

The test is neither obeyed by Laspeyres' Method nor by Paasche's Method.

Taking Paasche's Method, *omitting the factor 100*, we have,

$$I_{o,n} = \frac{\Sigma p_nq_n}{\Sigma p_oq_n} \text{ and } I_{n,o} = \frac{\Sigma p_oq_o}{\Sigma p_nq_o}$$

but, $I_{o,n} \times I_{n,o} = \frac{\Sigma p_nq_n}{\Sigma p_oq_n} \times \frac{\Sigma p_oq_o}{\Sigma p_nq_o}$ is not in general equal to 1.

Similarly taking Laspeyres' Method, *omitting the factor 100*, we have

$$I_{o,n}=\frac{\Sigma p_n q_o}{\Sigma p_o q_o} \text{ and } I_{n,o}=\frac{\Sigma p_o q_n}{\Sigma p_n q_n}$$

but, $I_{o,n}\times I_{n,o}=\frac{\Sigma p_n q_o}{\Sigma p_o q_o}\times\frac{\Sigma p_o q_n}{\Sigma p_n q_n}$ is not in general equal to 1.

The Fisher's Index Number satisfies this test. We have, *omitting the factor 100*,

$$I_{o,n}=\sqrt{\frac{\Sigma q_n p_n}{\Sigma q_n p_o}\times\frac{\Sigma q_o p_n}{\Sigma q_o p_o}} \text{ and } I_{n,o}=\sqrt{\frac{\Sigma p_o q_o}{\Sigma q_o p_n}\times\frac{\Sigma q_n p_o}{\Sigma q_n p_n}}$$

which gives, $I_{o,n}\times I_{n,o}=1$.

(2) *Factor Reversal Test.*

The product of the Price Index Number and the corresponding Quantity Index Number obtained by interchanging p by q and q by p should be the Value Index.

Neither the Laspeyres' Method nor the Paasche's Method satisfies this test.

This test is satisfied by Fisher's Index Number. For, *omitting the factor 100*, we have,

$$P_{o,n}=\sqrt{\frac{\Sigma p_n q_o}{\Sigma p_o q_o}\times\frac{\Sigma p_n q_n}{\Sigma p_o q_n}}$$

and $$Q_{o,n}=\sqrt{\frac{\Sigma q_n p_o}{\Sigma p_o q_o}\times\frac{\Sigma p_n q_n}{\Sigma q_o p_n}}$$

which gives, $$P_{o,n}\times Q_{o,n}=\sqrt{\frac{\Sigma p_n q_o}{\Sigma p_o q_o}\times\frac{\Sigma p_n q_n}{\Sigma p_o q_n}\times\frac{\Sigma q_n p_o}{\Sigma p_o q_o}\times\frac{\Sigma p_n q_n}{\Sigma q_o p_n}}$$

$$=\frac{\Sigma p_n q_n}{\Sigma p_o q_o}=V_{o,n}$$

where $P_{o,n}$, $Q_{o,n}$ and $V_{o,n}$ are respectivly the Price Index Number, Quantity Index Number and Value Index Number for time 'n' on time 'o' as base.

(3) *Circular Test.*

This is an extension of time reversal test. Symbolically,

$$I_{o,1}\times I_{1,2}\times I_{2,3}\times\cdots\times I_{(n-1),n}\times I_{n,o}=1.$$

This test is satisfied only by G. M. of relatives and by fixed weight aggregates.

EXAMPLE :

Using the following data, show that Laspeyres' Price Index formula does not satisfy the time reversal test.

Items	*Base year*		*Current year*	
	Price	*Quantity*	*Price*	*Quantity*
A	6	50	10	56
B	2	100	2	120
C	4	60	6	60
D	10	30	12	24
E	8	40	12	36

SOLUTION :

Let p_0, p_n denote prices in the base year and current year and q_0, q_n the quantities in the base year and current year.

Items	p_0	q_0	p_n	q_n	p_0q_0	p_nq_0	p_0q_n	p_nq_n
A	6	50	10	56	300	500	336	560
B	2	100	2	120	200	200	240	240
C	4	60	6	60	240	360	240	360
D	10	30	12	24	300	360	240	288
E	8	40	12	36	320	480	288	432

$\Sigma p_0q_0 = 1360$, $\Sigma p_nq_0 = 1900$, $\Sigma p_0q_n = 1344$, $\Sigma p_nq_n = 1880$

Omitting the factor 100 from each index,

$$\text{Laspeyres' Index Number } (I_{o,n}) = \frac{1900}{1360} = 1{\cdot}39$$

$$I_{n,o} = \frac{1344}{1880} = 0{\cdot}71.$$

$\therefore \quad I_{o,n} \times I_{n,o} = 1{\cdot}39 \times 0{\cdot}71 \neq 1.$

$\therefore$ Laspeyres' formula does not satisfy time reversal test.

EXAMPLE :

Using above data, show that Fisher's Ideal formula satisfies Factor Reversal Test.

SOLUTION :

Omitting the factor 100 from each index,

Fisher's Ideal Price Index Number $(P_{o,n}) = \sqrt{\frac{1900}{1360} \times \frac{1880}{1344}}$ and

Fisher's Ideal Quantity Index Number $(Q_{o,n}) = \sqrt{\frac{1344}{1360} \times \frac{1880}{1900}}$

$$\therefore \; P_{o,n} \times Q_{o,n} = \sqrt{\frac{1900}{1360} \times \frac{1880}{1344} \times \frac{1344}{1360} \times \frac{1880}{1900}}$$

$$= \frac{1880}{1360} = \frac{\Sigma p_n q_n}{\Sigma p_o q_o} = \text{Value Index.}$$

Thus Fisher's Ideal Index formula satisfies Factor Reversal Test.

Purchasing Power of Money.

It is well-known that the *value of money*, also called *purchasing power of money*, changes from time to time with the changes of price line. When price line is steady, the value of money is stable. It goes down with rising price line and vice versa.

If in a period of reference the Price Index has risen to, say 140, obviously what a rupee will buy is only $\frac{100}{140}$ or $\frac{5}{7}$th of what it used to buy in the base period. This means the money value of rupee one in the base period is dropped to $\frac{5}{7} = {\cdot}71$ or 71 paisa approximately, or we may say the purchasing power of a rupee with reference to a particular base period is now approximately 71 paisa. Thus it is seen that purchasing power of money or value of money is, really, inversely proportional to the price index and at a given period with reference to the base period may be measured in the following,

Purchasing power of money or value of money

$$= \frac{100}{\text{Price Index Number}}$$

Again, if the price index for the year, say, 1960 be 110·3 and the price index for the year, say, 1950 be 98·4, then the purchasing power of a Rupee of 1950 will be (110·3 ÷ 98·4) of 1960.

Thus the purchasing power of a Rupee in any year in terms of 1960-Rupee will be given by,

$$\text{Purchasing Power} = \frac{\text{Price Index Number for 1960}}{\text{Price Index Number for the year}}.$$

The index number also provides an excellent material for transforming the money (nominal) income to real income. Real income is the equivalent income in terms of the value of money in the base year. It is obtained by dividing the money income by an *appropriate* Price Index. This process is known as 'statistical deflation'.

$$\text{Thus, Real Income or Wage} = \frac{\text{Money Income or Wage}}{\text{Price Index Number}} \times 100$$

and the Real Wage or Income Index Number

$$= \frac{\text{Real Wage of current year}}{\text{Real Wage of base year}} \times 100,$$

$$= \frac{\text{Index of Money Wages}}{\text{Price Index Number}} \times 100.$$

EXAMPLE :

The table below gives the average wages in Rs. per day of a group of workers from 1947 to 1951 and the consumer price index for these years. Determine the real wages of the workers.

Year	1947	1948	1949	1950	1951
Average Wages	1·19	1·33	1·44	1·57	1·75
Consumer Price Index No. 1947—49 = 100	95·5	102·8	101·8	102·8	111·0

SOLUTION :

$$\text{Real wage} = \frac{\text{Actual wage}}{\text{Consumer price index number}} \times 100$$

Calculation of Real Wages

Year	Average Wages	Consumer Price Index No.	Real Wages (Rs.)
1947	1·19	95·5	(1·19 ÷ 95·5) × 100 = 1·25
1948	1·33	102·8	(1·33 ÷ 102·8) × 100 = 1·29
1949	1·44	101·8	(1·44 ÷ 101·8) × 100 = 1·41
1950	1·57	102·8	(1·57 ÷ 102·8) × 100 = 1·53
1951	1·75	111·0	(1·75 ÷ 111·0) × 100 = 1·58

Uses of Index Numbers.

Price Index Numbers guide the businessmen in forming suitable policies. Stable price line or slowly rising price line is considered favourable to the stability of business. When the price line is high businessmen frame one policy and when low they frame a different one. Price index numbers measure the changes in the price line in order to study the changes and try to control them.

Wages and salaries are being adjusted now-a-days on the basis of appropriate consumer price index numbers.

Price index is also used as a reliable measure of the purchasing power of money (value of money) which varies inversely with it. Money income is often transformed into real income by dividing it by an appropriate price index.

Index of industrial production is used to study the general progress of the business condition of the country and to compare the production of one's own business.

Investment index numbers guide the economists, speculators and bankers in various ways. Index numbers can also be used for measuring changes over space, viz. Consumers price index of textile workers in two cities for the same period of time.

It has also application in measuring the difference in the level of intelligence among students in different institutions, in standardizing the birth and death rates, to measure changes in the effectiveness of school systems. They are also used to forecast the future trend.

Index numbers are most widely used in the evaluation of general economic behaviour in the country. In the economy as a whole the index numbers of overall business activity are called 'business barometers'.

"*Index numbers are to-day one of the most widely used statistical devices....... They are used to take the pulse of the economy and they have come to be used as indication of inflationary or deflationary tendencies.*" —G. Simpson and F. Kafka.

Construction of Wholesale Price Index Number.

Wholesale price index number is intended to measure relative variation in the general price level in a given period of time compared to some base period. To study the relative change in the general price level, data about the wholesale prices of the commodities marketed are to be collected. Since it is not possible to include all the commodities in the index, a sample of the commodities has to be taken in such a way that they will be manageable in number and representative of the taste, habits and customs of the people. The items included should be of standardized quality and are broadly divided into five major groups :

(1) *Food* ;

(2) *Fuel, Light, Power and Lubricants* ;

(3) *Liquor and Tobacco* ;

(4) *Industrial Raw Material* ; and

(5) *Manufactures*—(*a*) *Intermediate Products* ; (*b*) *Finished Products.*

Each group is again divided into a number of sub-groups. For example, the group 'food' is subdivided in three subgroups, *i.e.*, (i) *Cereals*, (ii) *Pulses* and (iii) *Others*. Weights are, in general, proportionate to the value of the quantities marketed in the base period. A period of economic stability should be taken as base period. Wholesale prices of the commodities included in the index are to be collected at regular intervals of time from Govt. agencies and commercial centres. Price quotations are taken generally once a week and index number is calculated. The G. M. of weekly indices is taken as the index number for a month.

The method for constructing the index number is Weighted Arithmetic Mean of Price Relatives. Geometric Mean of Price Relatives can also be used in the calculation of the index number.

Index of Industrial Production

Index of industrial production is generally intended to measure the relative variation in the level of industrial production in a country in a given period of time compared to some base period. It is constructed to study the changes in the *quantum* of production and not in *values*. For the compilation of such index numbers, data about the level of industrial output of various industries are to be collected. Since it is not possible to include all the items of industrial production in the index number, a sample of items has to be taken in such a way that they are manageable in number and representative in character. The items included should be of standard quality and are broadly divided into six major classes :

(1) *Textile Industries* ;

(2) *Mining Industries* ;

(3) *Metallurgical Industries* ;

(4) *Mechanical Industries* ;

(5) *Industries subject to excise duties* ; and

(6) *Miscellaneous.*

Each class is again divided into a number of important items. For example, the class textile industries include cotton, woollen, silk, etc.

The method used for constructing such index numbers is, in general, weighted arithmetic mean of production relatives. Weighted geometric mean of production relatives may also be used in the construction of this index number.

Weights are selected on the basis of relative importance of different industries. Usually the weights are based on the values of net output of different industries during the base period.

Limitations of Index Numbers.

(1) Index numbers are, in general, based on samples. Items and quantities are obtained on the basis of sampling and since sampling is always subjected to bias, hence errors are introduced. So efforts must be made to minimise the errors.

(2) With the passage of time the tastes and habits of the people change and as such there is always a risk of a change in the qualities of the items. Introduction of items of new quality and the obsolescence of others makes comparison over long period unreliable.

(3) Though different methods of construction of index numbers will give different results, sometime the difference is substantial still all the indices point to the same direction and the trends generally agree unless there are rapid changes in conditions.

(4) Index numbers can also be manipulated to suit ones purpose. Base period having abnormally high profit may be taken to show the current profit very low. To show current prices very high, base period having abnormally low prices may be chosen sometimes.

(5) Data collected from different sources and regions may not always be reliable since the quality, honesty and intelligence of all the investigators are never the same. This with non-availability of a perfectly normal base period possesses a serious limitation of the index number.

Miscellaneous Examples.

1. In 1976 the average price of a commodity was 20% more than in 1975, but 20% less than in 1974 and it was 50% more than in 1977. Reduce the data to price relatives :

(i) Using 1975 as base ; and

(ii) With 1974-75 as base average. [I. C. W. A. Dec. 1978]

SOLUTION :

Let the price of 1976 = 100

So, the price of 1975 $= \frac{100 \times 100}{120} = 83{\cdot}33$, the price of 1974 is

$= \frac{100 \times 100}{80} = 125$ and the price of 1977 $= \frac{100 \times 100}{150} = 66{\cdot}67$.

Average price for two years 1974 & 1975 $= \frac{125 + 83{\cdot}33}{2} = 104{\cdot}16$

Calculation of Price Relatives

Year	Price	P.R. with 1975 as base	P.R. with 1974-75 as base
1974	125	$\frac{100\times125}{83\cdot33}=150$	$\frac{100\times125}{104\cdot16}=120$
1975	83·33	100	$\frac{100\times83\cdot33}{104\cdot16}=80$
1976	100	$\frac{100\times100}{83\cdot33}=120$	$\frac{100\times100}{104\cdot16}=96$
1977	66·67	$\frac{100\times66\cdot67}{83\cdot33}=80$	$\frac{100\times66\cdot67}{104\cdot16}=64$

2. On the basis of the following data, compute the Index Number (i) using simple average of price relatives, and (ii) with the help of geometric mean of price relatives :

Commodites	A	B	C	D	E
Prices (in Rs.) 1948	10	18	12	20	25
Prices (in Rs.) 1950	12	16	10	8	8

SOLUTION :

Let p_o and p_1 be the prices in 1948 and 1950 respectively.

Calculation of Index Numbers

Commodities	p_o	p_1	P.R. $=\frac{p_1}{p_o}\times100$	$\log\frac{p_1}{p_o}\times100$
A	10	12	120·00	2·0792
B	18	16	88·89	1·9488
C	12	10	83·33	1·9208
D	20	8	40·00	1·6021
E	25	8	32·00	1·5051

$$\sum\frac{p_1}{p_o}\times100=364\cdot22,\quad \sum\log\frac{p_1}{p_o}\times100=9\cdot0560$$

Index Number (using simple average of P.R.) $=\frac{364\cdot22}{5}=72\cdot84$

$$\text{Index Number (using G.M. of P.R.)} = \text{Antilog}\left(\frac{\Sigma \log \frac{p_1}{p_o} \times 100}{n}\right)$$

$$= \text{Antilog}\ \frac{9{\cdot}0560}{5}$$

$$= \text{Antilog}\ 1{\cdot}8112$$

$$= 64{\cdot}74.$$

3. Construct Fisher's Ideal Index Number for the following data :

Commodity	1960 (Base Year)		1968 (Current Year)	
	Price	*Quantity*	*Price*	*Quantity*
A	8	6	12	5
B	10	5	11	6
C	7	8	8	5

[I. C. W. A. 1970]

SOLUTION :

Let p_o and p_1 represent prices in 1960 and 1961 and q_o and q_1 be the quantities in 1960 and 1961 respectively.

Computation of Fisher's Ideal Index Number

	1960		1968					
Commodity	p_o	q_o	p_1	q_1	p_oq_o	p_1q_o	p_oq_1	p_1q_1
A	8	6	12	5	48	72	40	60
B	10	5	11	6	50	55	60	66
C	7	8	8	5	56	64	35	40

$\Sigma p_oq_o = 154$, $\Sigma p_1q_o = 191$, $\Sigma p_oq_1 = 135$, $\Sigma p_1q_1 = 166$.

$$\text{Fisher's Ideal Index Number} = \sqrt{\frac{\Sigma p_1q_o}{\Sigma p_oq_o} \times \frac{\Sigma p_1q_1}{\Sigma p_oq_1}} \times 100$$

$$= \sqrt{\frac{191}{154} \times \frac{166}{135}} \times 100$$

$$= 123{\cdot}5.$$

4. The price of agricultural commodities for 1966-67 and for the month of December, 1970 are given below along with the value of these commodities in 1966-67 :

Commodities	*Unit*	*Price*		Value of output in million rupees
		1966-67	Dec. 70	
Rice	Md.	13·75	13·75	8364
Wheat	Md.	9·70	9·70	2207
Jowar	Md.	6·03	8·00	876
Cotton (raw)	784 lb.	466·00	433·00	701
Tea	lb.	1·25	1·75	534

[Delhi Univ. 1975]

Calculate the Weighted Index Number of prices of these commodities for Dec. 1970, taking 1966-67 as base.

SOLUTION :

Since the *base period* values of output are given the Index Number may be calculated using Weighted A.M. of price relatives, weights being the values of output.

Let p_0 and p_1 represents the prices in 1966-67 and Dec. 1970 respectively and w be the value of output.

Computation of Weighted Index Number

Commodities	*Unit*	p_0	p_1	w	$P.R. = \frac{p_1}{p_0} \times 100$	$P.R. \times w$
Rice	Md.	13·75	13·75	8364	100·00	8,36,400·00
Wheat	Md.	9·70	9·70	2207	100·00	2,20,700·00
Jowar	Md.	6·03	8·00	876	132·67	1,16,218·92
Cotton (raw)	784 lb.	466·00	433·00	701	92·92	65,136·92
Tea	lb.	1·25	1·75	534	140·00	74,760·00

Σ (price relative × weight) = 13,13,215·84 ; $\Sigma w = 12,682$

$$\text{Weighted Index Number} = \frac{1313215·84}{12682} = 103·55.$$

5. From the following data, calculate the Wholesale Price Index Number for the 5 groups combined. [C. U., M. Com. 1970]

Group	*Weight*	*Index Number* for week-ending 27. 9. 69 (Base : 1952-53 = 100)
Food articles	50	241
Liquor and tobacco	2	221
Fuel, power, etc.	3	204
Industrial raw materials	16	256
Manufactured commodities	29	179

SOLUTION :

Let W and I represent the Weight and Index Number for the week-ending 27. 9. 69 respectively.

Compilation of Wholesale Price Index Number

Group	W	I	IW
Food articles	50	241	12,050
Liquor and tobacco	2	221	442
Fuel, power, etc.	3	204	612
Industrial raw materials	16	256	4,096
Manufactured commodities	29	179	5,191

$$\Sigma W = 100 \; ; \quad \Sigma IW = 22{,}391$$

$$\therefore \text{ Wholesale Price Index Number} = \frac{\Sigma IW}{\Sigma W} = \frac{22{,}391}{100} = 223{\cdot}91.$$

6. The data below show the percentage increases in price of a few selected food items and the weights attached to each of them. Calculate the Index Number for the food group.

Food items :	*Rice*	*Wheat*	*Dal*	*Ghee*	*Oil*	*Spices*	*Milk*	*Fish*	*Vegetable*	*Refreshments*
Weight :	33	11	8	5	5	3	7	9	9	10
Percentage increase in price :	180	202	115	212	175	517	260	426	332	279

Using the above food index and the information given below, calculate the cost of living index number.

Group	Food	Clothing	Fuel and Light	Rent	Miscellaneous
Index	—	310	220	150	300
Weight	60	5	8	9	18

[I. C. W. A.'72]

SOLUTION :

Calculation for Food Index Number

Food items	*Weight (w)*	*Percentage increase in price (i)*	*Current Index (I)*	*Iw*
Rice	33	180	280	9240
Wheat	11	202	302	3322
Dal	8	115	215	1720
Ghee	5	212	312	1560
Oil	5	175	275	1375
Spices	3	517	617	1851
Milk	7	260	360	2520
Fish	9	426	526	4734
Vegetable	9	332	432	3888
Refreshments	10	279	379	3790

Note : Current Index = Percentage increase + 100

$$\Sigma w = 100 \; ; \; \Sigma Iw = 34000$$

$$\therefore \quad \text{Food Index Number} = \frac{\Sigma Iw}{\Sigma w} = \frac{34000}{100} = 340.$$

Calculations for C. L. I.

Group	*Index (I)*	*Weight (w)*	*Iw*
Food	340	60	20,400
Clothing	310	5	1,550
Fuel and light	220	8	1,760
Rent	150	9	1,350
Miscellaneous	300	18	5,400

$\therefore \quad \Sigma w = 100 \; ; \; \Sigma Iw = 30{,}460.$

$\therefore$ Cost of Living Index Number $= \dfrac{30{,}460}{100} = 304·6.$

7. The following table gives the annual income of a person and the general price index number for five years :

Year :	1970	1971	1972	1973	1974
Income (Rs.) :	3600	4200	5000	5500	6000
General Price Index Number :	100	104	115	160	280

Determine the real income of the person.

SOLUTION :

$$\text{Real Income} = \frac{\text{Actual Income}}{\text{Price Index}} \times 100.$$

Calculation of Real Income

Year	*Income* (Rs.)	*Index Number*	*Real Income* (Rs.)
1970	3600	100	$\frac{3600}{100} \times 100 = 3600·00$
1971	4200	104	$\frac{4200}{104} \times 100 = 4038·46$
1972	5000	115	$\frac{5000}{115} \times 100 = 4347·83$
1973	5500	160	$\frac{5500}{160} \times 100 = 3437·50$
1974	6000	280	$\frac{6000}{280} \times 100 = 2142·85$

8. The index numbers of wholesale prices for the years 1947—54 with 1947—49 = 100 are given below. Determine the wholesale purchasing power of a Rupee in terms of 1954-Rupee in each of the given years.

Year	1947	1948	1949	1950	1951	1952	1953	1954
Wholesale price index (1947—49 = 100)	96·4	104·4	99·2	103·1	114·8	111·6	110·1	110·3

[C. U. 1963]

SOLUTION :

We have, purchasing power of a rupee $= \dfrac{\text{Index Number for 1954}}{\text{Index Number for the year}}$.

Calculation for Purchasing Power of a Rupee

Year	*Wholesale Price Index*	*Purchasing Power of a Rupee* (in terms of 1954-Rupee)
1947	96·4	110·3 ÷ 96·4 = 1·14
1948	104·4	110·3 ÷ 104·4 = 1·06
1949	99·2	110·3 ÷ 99·2 = 1·11
1950	103·1	110·3 ÷ 103·1 = 1·07
1951	114·8	110·3 ÷ 114·8 = 0·96
1952	111·6	110·3 ÷ 111·6 = 0·99
1953	110·1	110·3 ÷ 110·1 = 1·00
1954	110·3	110·3 ÷ 110·3 = 1·00

9. Calculate the index of industrial production of the following, using weighted A.M. of production relatives.

Items	*Production*		*Value of Net output* (Rs.)
	1970	1980	
A	1,00,000 litres	1,60,000 litres	3,50,000
B	1,20,000 kg.	1,50,000 kg.	2,00,000
C	5,000 m. tons.	7000 m. tons.	9,00,000

SOLUTION :

Items	*Production Relatives*
A	(160000 ÷ 100000) × 100 = 160
B	(150000 ÷ 120000) × 100 = 125
C	(7000 ÷ 5000) × 100 = 140

∴ Index of Industrial Production

$$= \frac{160 \times 350000 + 125 \times 200000 + 140 \times 900000}{350000 + 200000 + 900000}$$

$= 143.$

EXERCISE 11

1. The average price of mustard oil per quintal in the years 1954 to 1958 are given below :

Year :	1954	1955	1956	1957	1958
Price :	295	275	300	225	250

Find the index numbers for all the years taking 1956 as the base year. (B. U. 1964) [*Ans.* 98·3, 91·7, 100, 75, 83·3]

2. Find the Index Number by method of relatives (using A M.) from the following data :

Commodity	*Base Price*	*Current Price*
Rice	35	42
Wheat	30	35
Pulse	40	38
Fish	105	120

(C. U. 1973) [*Ans.* 107·5]

3. The following table gives the average wholesale prices in rupees for three commodities during the years 1965—70. Construct the index numbers for all the years by taking 1965 as the base year.

Commodity	1965	1966	1967	1968	1969	1970
A	25·3	30·8	33·4	35·5	35·3	36·0
B	17·3	14·5	4·9	5·7	17·1	11·6
C	7·8	5·4	6·7	5·6	7 2	10·2

[*Ans.* 100, 92, 82, 82, 111, 113]

4. Find by A.M. method, the index number from the following data :

Commodity	*Weight*	*Current Price*	*Base Price*
Cloth	13	250	225
Wheat	18	26	22
Rice	25	32	26
Potato	8	65	70

(B. U. 1970) [*Ans.* 115·5]

5. Given below are the data on prices of some consumer goods and the weights attached to the various items. Compute price index numbers for the year 1969 (Base 1968=100), using (i) simple average, and (ii) weighted average of price relatives.

Items	*Unit*	Price (Rs.) 1968	 1969	*Weight*
Wheat	Kg.	0·50	0·75	2
Milk	Litre	0·60	0·75	5
Egg	Dozen	2·00	2·40	4
Sugar	Kg.	1·80	2·10	8
Shoes	Pair	8·00	10·00	1

(I. C. W. A. 1979) [(i) *Ans.* 127·4 ; (ii) 123·3]

6. Find the weighted index number, using the following data :

Items	*Index*	*Weight*
Food	152	48
Clothing	110	5
Rent	130	10
Fuel and lighting	100	12
Miscellaneous	80	15

[*Ans.* 128·29]

7. Apply Fisher's method and calculate the index number for 1974 with 1973 as base year from the following data :

	1973		1974	
Items	*Price*	*Quantity*	*Price*	*Quantity*
A	10	4	12	3
B	15	6	20	5
C	2	5	5	6
D	4	4	4	4

[*Ans.* 139·9]

8. Construct a suitable index number with the help of the following data with 1965 as base :

Commodity :	Wheat		Rice		Gram	
Year	*Price*	*Quantity*	*Price*	*Quantity*	*Price*	*Quantity*
1965	14	15	20	5	4	10
1969	24	12	27	4	7	8

[*Ans.* Fisher's Index = 161·4]

9. Calculate Laspeyres' and Paasche's Index Numbers from the following data :

	Base year		Current year	
	Quantity	*Price per* ℔. (Rs.)	*Quantity*	*Price per* ℔. (Rs.)
Rice	6·0	·40	7·0	·30
Meat	4·0	·45	5·0	·50
Tea	0·5	·90	1·5	·40

[*Ans.* Laspeyres' Index = 86·02, Paasche's Index = 81·25]

10. Construct index number of price from the following data by applying (i) Laspeyres' Method, and (ii) Paasche's Method.

	Base year		Current year	
Commodities	*Price*	*Quantity*	*Price*	*Quantity*
A	2	8	4	6
B	5	10	6	5
C	4	14	5	10
D	2	19	2	13

[*Ans.* Laspeyres' Index = 125, Paasche's Index = 126·2]

11. In the construction of rural price index number in a certain centre following results are obtained :

		Group Index Number	
Group	*Weight*	Oct. 1966	Nov. 1966
Food	81	323	344
Lighting	2	190	186
Clothing	8	432	397
Miscellaneous	9	369	377

Find the rural price index numbers for the two months.

[*Ans.* For Oct. 333·2 ; For Nov. 348·5]

12. Using the data given below calculate price index numbers for the year 1958 by (i) Laspeyres' Method, (ii) Paasche's Method, and (iii) Fisher's Method with the year 1949 as base.

Commodity	*Price* (Rs.)		*Quantity* ('000 Kg.)	
	1949	1958	1949	1958
Rice	9·3	4·5	100	90
Wheat	6·4	3·7	11	10
Pulse	5·1	2·7	5	3

[*Ans.* 49·1, 49·1, 49·1]

13. Calculate the cost of living index number from the following data :

Items	Price *Base year*	 *Current year*	*Weights*
Food	30	47	4
Fuel	8	12	1
Clothing	14	18	3
Rent	22	15	2
Miscellaneous	25	30	1

[*Ans.* 115·84]

14. In the construction of a certain C. L. I. number, the following group index numbers are found. Calculate C. L. I. by using (i) the Weighted A.M. and (ii) the Weighted G.M.

Group	*Index Number*	*Weight*
Food	350	5
Fuel and lighting	200	1
Clothing	240	1
House rent	160	1
Miscellaneous	250	2

[*Ans.* C.L.I. using Weighted A.M. = 285
C.L.I. using Weighted G.M. = 275·4]

15. Find the index number for the years 1967, 1968 and 1969 by the chain-base method, with base year 1966, from the following :

Year :	1966	1967	1968	1969
Link Index :	100	110	95·5	109·5

(I. C. W. A. 1969)

[*Ans.* Chain indices, 1967 = 110, 1968 = 105·05, 1969 = 115·03]

16. From the following table calculate the average percentage increase in cost in 1945 over 1939 of manufacturing thin grade of linoleum to which the data refer :

Item of cost	*Linoleum Production costs*	
	Item as % of total cost	% increase in cost of items in 1945 over 1939
Materials	48	97
Direct Labour	15	43
Manufacturing on costs	21	114
Administrative costs	16	10

(C. U. 1966) [*Ans.* 78·55]

17. From the following table calculate Paasche's quantity index number for 1969 with 1951 as base :

Commodity	*Quantity*		*Value*
	1951	1969	1969
A	54	250	540
B	93	75	825
C	18	56	448
D	6	8	56
E	23	47	141

(I. C. W. A. 1972) [*Ans.* 144]

18. Construct Weighted Aggregative Quantity Index Numbers taking 1976 as the base :

Commodity	*Average price per unit* (Rs.)	*Production* ('0000 tons) 1976	1977	1978
A	2·00	150	180	192
B	3·00	240	220	160
C	0·50	1100	1200	1500
D	4·50	22	24	27

[*Ans.* 103·53, 103·98]

19. From the following data, prepare quantity index numbers for the year 1974 taking 1968 as the base year :

Year	*Commodity I*		*Commodity II*		*Commodity III*	
	Price	*Quantity*	*Price*	*Quantity*	*Price*	*Quantity*
1968	5	10	8	6	6	3
1974	4	12	7	7	5	4

(I. C. W. A. 1975) [*Ans.* Laspeyres' Index = 120·68 ; Paasche's Index = 120·62 ; Fisher's Index = 120·65]

20. The average weekly wages for all manufacturing industries for a number of months in 1960 are 78·52, 79 71, 78·55, 78·17, 78·99 ; the corresponding consumer price index numbers are 115, 114·7, 114·6, 114·6, 114·9. Find the real wages for the different months and calculate the percentage change in the real wages during the period. (C. U. 1968)

[*Ans.* 68·28, 69·49, 68·54, 68·21, 68·75 and 0·69%]

21. During a certain period the C. L. I. goes up from 110 to 200 and the salary of a worker is also raised from Rs. 325 to Rs. 500. Does the worker really gain, and if so, by how much in real terms ?

[*Ans.* No. Real Wage decreases by Rs. 45]

22. From the following prove that the Fisher's Ideal Index Number satisfies both the Time Reversal Test and Factor Reversal Test.

	Base year		Current year	
Commodity	*Price*	*Quantity*	*Price*	*Quantity*
A	6	50	10	56
B	2	100	2	120
C	4	60	6	60
D	10	30	12	24

23. In a working-class consumer price index number of a particular town the weights corresponding to different groups of items were as follows :

Food—55, Fuel—15, Clothing—10, Rent—8, and Miscellaneous—12.

In Oct., 1972, the D. A. was fixed by a Mill of that town at 182 percent of the workers which fully compensated for the rise in prices of food and rent but did not compensate for anything else. Another Mill of the same town paid D. A. of 46'5 percent which compensated for the rise in fuel and miscellaneous groups. It is known that rise in food is double that of fuel and the rise in miscellaneous group is double that of rent.

Find the rise of food, fuel, rent and miscellaneous groups.

[Rise of food—317'2, Fuel—158'6, Rent—94'5, Miscellaneous—189'0 ; assuming no rise in clothing].

24. Following table shows the annual wages of a labourer and the price index numbers. Prepare the real wage index numbers for the labourer :

Year :	1967	1968	1969	1970	1971	1972	1973
Wages in Rs. :	200	240	350	360	360	370	375
Price Index :	100	160	280	290	300	320	330

[*Ans.* 100, 75, 62'5, 62, 60, 57'8, 56'8]

25. Monthly wages average in different years is as follows :

Year :	1967	1968	1969	1970	1971	1972	1973
Wages (Rs.) :	200	240	350	360	360	380	400
Price Index :	100	150	200	220	230	250	250

Calculate real wages index numbers.

(I. C. W. A. 1979) [*Ans.* 100, 80, 87'5, 82, 78'5, 76, 80]

12

TIME SERIES

Introduction.

A time series is defined as data arranged chronologically. It is a set of observations taken, usually, at equal intervals of time on a variable which is dependent on time. Such a series of observations discloses the changes or variations in the value of the variable due to changes in time. The sale of a commodity in different years, quarterly export of tea, monthly production of cement, rainfall in centimeter on different days, number of accidents on various days, etc., all give rise to time series data.

The time series data are playing increasingly significant role in all kinds of economic, business and commercial activities. It enables the economists and businessmen to know the changes that have taken place in the past and to compare the present activities with those for another period. It also enables them to know which products appear to be gaining or losing ground. It also helps them in predicting, to a certain limit, the future behaviour of the economic and business activities which is regarded most important to them before they proceed with their planning and budgeting. Modern statistical techniques enable us to know the factors responsible for the development in economic and business activities in the past and the way the same factors will operate in future. Once the regularity of occurrence of such factors over a sufficiently long period is established, the prediction of probable future variation, within certain limits, is possible.

Characteristics of a Time Series.

A time series data may best be studied by plotting them on a graph paper. When plotted it will be seen that most of the time series data show :

(i) some movements exhibiting persistent growth or decline,

(ii) some movements regular and periodic in nature with period not more than one year,

(iii) some fairly regular and periodic with period of duration of more than a year, and finally,

(iv) some irregular, mild or violent movements.

However, all the series may not show all the movements stated above and some series may show some type of movement not mentioned above.

Thus a time series, in general, is the result of *four types of movements.* They are as follows :

(1) *A Secular (basic) Trend.*

This is a smooth, regular and long-term upward or downward movement in the data. It reveals the general tendency of the data. Almost all the time series show a tendency to increase or to decrease gradually over a long period of time. The increase may be due to growth of population together with improvement of purchasing power of people, advancement in technology, scientific management, personnel management, quality control and many other specific factors. A time series may show a downward trend due to availability of better and cheaper substitute or difficulty in obtaining raw materials or decrease in the demand of the product, etc. Some series, however, may initially show a period of steady growth, reverse themselves and then show a period of decline and vice versa.

(2). *Seasonal Variation.*

This is a short-term periodic movement whose period is not longer than a year. It is uniform and regular in nature. This short-term movement is mainly due to the climatic changes or to holidays or to social customs, trading and other habits of the people, etc. Most of the business activities have slack periods and brisk periods every year and these slack periods and brisk periods are mostly related to the changing seasons (weather) or to holidays or to social customs, trading and other habits of the people and repeat with remarkable regularity after a period of time which is not longer than a year. The upward and the downward movement of the data characterising the slack periods and brisk periods of business activity which is being repeated with remarkable regularity with a period of not more than a year is conveniently termed as seasonal variation. Sale of woollen goods increases during winter and falls during summer, cold drinks have a greater sale in the summer, sale of garments are heavy during Pujas, are all examples of seasonal variation with periods of 12 months. But in many cases the period may be weekly, quarterly or monthly. For example, the withdrawal from banks are heavy on the first day of the month, the traffic and telephone calls have a peak during certain hours of the day, the number of books borrowed from a library has peak periods during some days of a week, etc.

(3) *Cyclical Variation.*

These are oscillatory movements with a period more than a year. Such movements do not ordinarily exhibit regular periodicity. In economic and business series they correspond to the business cycle. Almost all business and economic activities have *four* distinct phases : *Prosperity, Decline, Depression* and *Recovery* ; they constitute the business cycle and these four phases recur after a period of time which is longer than a year with rough regularity. These four phases are generated by factors other than changes in climate, social customs and those which create seasonal variations. Each phase of the cycle changes into the phase which follows it. Prosperity is followed by Decline until Depression is reached. Then the Depression is followed by Recovery and back to Prosperity. This type of upward and downward movements characterising the period of prosperity and period of depression of business activity which is being repeated with rough regularity with period more than a year is termed as cyclical variation.

(4) *Irregular Variation.*

These variations are of two types—*Catast ophic* and *Accidental* (residual). Catastrophic variations are due to specific events such as strikes, fires, earthquakes, floods, etc. The accidental variations are due to multiplicity of causes of unknown origin. They are minor variations and too small to merit individual consideration. These variations may be of random nature. There are variations which are left after all other variations have been accounted for.

The analysis of time series consists in separating the four constituent parts of the series, namely, *the trend, the cyclical variation, the seasonal variation,* and *the irregular variations,* analysing each constituent parts separately and then recombine the series in order to describe the observed variations in the variable of interest.

The Classical Model.

In classical time series analysis it is assumed that there is either a multiplicative or an additive relationship between the four components of the series, that is, it is assumed that any particular value of the data is either the resultant of the product of individual components or the resultant of the sum of the individual components.

Symbolically,

$$Y = T \times C \times S \times I \quad \text{(Multiplicative model)}$$

$$\text{or} \quad Y = T + C + S + I \quad \text{(Additive model)}$$

Where Y is the original data, T is the trend component, S is the seasonal component, C is the cyclical component and I is the irregular component.

The additive model, now-a-days, is not used in practice as it does not suit most of the economic time series. The multiplicative model is widely used as it portrays actual experience more closely than that of additive model. But the ultimate criterion for a given situation is to use the model that fits the data best.

Measurement of Trend.

Four methods are commonly used for measuring the trend :

(1) *Graphic Method* ;

(2) *Semi-Average Method* ;

(3) *Moving Average Method* ;

(4) *Method of Least Square.*

(1) *Graphic Method.*

In this method the series is plotted on a graph paper against time and a free-hand smooth curve is drawn by inspection between the points in such a way that the fluctuations in one direction are approximately equal to those in the other direction. This curve will show the long-term general tendency of the data, that is, the trend. As the time series data are shown along the vertical axis, the vertical distance of this curve estimates the trend value for each time period. This method is very simple but highly subjective. Because of its subjectiveness this method should not in general be used by inexperienced persons. However, it has considerable merit in the hand of an experienced person and is widely used in applied situations. By this method a quick estimate of the trend—both linear and non-linear—is obtained. It is always desirable to draw a graph to obtain a preliminary knowledge of the nature of trend in the time series data with a view to take a decision as to which type of mathematical trend will be appropriate.

(2) *Semi-Average Method.*

This method is applied when the trend is linear. In this method the whole time series data are divided into two equal parts and averages of each part are calculated. When the number of observations is even, there will be exactly two equal parts but when the number of observations is odd, the central value is generally neglected when the division is made. The two averages are then plotted at the mid-points of their respective time intervals and through these two points a straight line is drawn. This line will be the required trend line, the trend value for any time is obtained from the ordinate drawn at that point.

This method can also be applied when the increase or decrease is by constant ratio. In this case, however, instead of original data their logarithms are used as the basis of the calculation. The trend values at any time will then be the antilogarithm of the value of the ordinate at that point.

(3) *Moving Average Method*

The simplest method of smoothing out fluctuations and obtaining the trend values with fair degree of accuracy is the moving average method.

Moving averages are a number of arithmetic averages calculated from the time series data, each based on a *fixed number* of consecutive observations. If the time series data are yearly, then moving averages of period, say, r years are a series of arithmetic averages each of r consecutive observations. The first moving average is the average of first r observations and is placed at the time point midway between the time points of the first and the r-th observations of the series. For second moving average, drop the first observation and include the $(r+1)th$ observation in the calculation of the average. The second moving average is the average of second to $(r+1)th$ observations of the series and is placed at the middle of the period covering second to $(r+1)$ years. Similarly, the third moving average is the average of third to $(r+2)th$ observations of the series and is placed at the mid-point of the time interval covering third to $(r+2)$ years, and so on.

Since each moving average is placed at the time point midway between the time points of the first and the last observations included in the calculation of average, the moving average values do not correspond to any of the original periods when there are even number of periods, and hence two-item moving averages of the moving averages already obtained, have to be calculated to correspond them to any of the original periods. This process is called *recentering*.

The purpose of the moving average method is to smooth out cyclical, seasonal and irregular variations of the time series data in order to isolate the trend. It is observed that moving average will completely eliminate a fluctuation if the period of moving average be equal to the period of the fluctuation or its integral multiple.

When a series of yearly figures are given, the seasonal fluctuations, whose period is, generally, a year is automatically excluded from the series. The other fluctuation to be removed now is the cyclical fluctuation. If the period of the cyclical fluctuation is known, this can be eliminated by calculating moving averages taking the period of moving average equal to or an integral multiple of the period over which the cyclical fluctuations occur. When the period of cyclical fluctuation is not obvious then a graph of the actual data is to be drawn and the distance between two 'peaks' or two 'depressions' of the graph will be taken as period when the cycle is regular, and this period or an integral multiple of this period will be taken as the period of the moving average to smooth out the cyclical fluctuations. When the period of the cycle is not uniform, the average duration of the cycles or an integral multiple of it may be taken as the period.

When the *monthly or quarterly* figures are given then a twelve month or four quarter moving average is called for to smooth out seasonal fluctuations. Now if the cyclical fluctuation has a period of, say, four years then the moving average with a period of 48 months or 16 quarters will smooth out both the seasonal and the cyclical fluctuations. But if the period of the cyclical fluctuation be, say, 30 months then the moving average with a period of 60 months (the least common multiple of both the period of seasonal fluctuation and the period of cyclical fluctuation) is required to smooth out both the seasonal and the cyclical fluctuations.

Moving average, in general, cannot eliminate irregular fluctuations but it only reduces them.

Thus the moving averages with period same as the period of the cycle or its integral multiple will smooth out the seasonal and the cyclical fluctuations and give an estimate of the trend.

The moving average values, sometimes, do not follow the data which describe a curve unless some weighting schemes are used. The usual type of weighting is, however, binomial, which employ binomial coefficients as weights. There are other systems of weighting also. The main point in favour of weighted moving average is that they are both relatively smooth and sufficiently sensitive.

Merits and Demerits

(1) This method is flexible and not subjective. It is simple to understand and easy to adopt.

(2) This method is appropriate only when the trend is linear. If the trend is not linear, the moving average will over-estimate or under-estimate the trend value.

(3) Cyclical fluctuation may be eradicated completely if the cycles are regular and the period of moving average be equal to or an integral multiple of the period of fluctuation. In all other cases moving averages will reduce them.

(4) Trend values cannot be determined for some periods at the beginning and at the end.

(5) The method cannot be used for forecasting future trend as the moving averages assume no definite mathematical law of change.

(6) This method is very sensitive to a few very high and low values which the series may contain.

Example :

Compute five-yearly moving averages from the following :

Year :	1924	1925	1926	1927	1928	1929	1930	1931	1932
Annual Sales :	6·4	4·3	4·3	3·4	4·4	5·4	3·4	2·4	1·4

SOLUTION :

Calculations for 5-yearly Moving Averages

Year	Annual Sales	5-year Moving Total	5-year Moving Average
1924	6·4	—	—
1925	4·3	—	—
1926	4·3	22·8	4·56
1927	3·4	21·8	4·36
1928	4·4	20·9	4·18
1929	5·4	19·0	3·80
1930	3·4	16·0	3·20
1931	2·4	—	—
1932	1·4	—	—

[**Working Notes.**

First moving total $= 6·4 + 4·3 + 4·3 + 3·4 + 4·4 = 22·8$

Second moving total $= 4·3 + 4·3 + 3·4 + 4·4 + 5·4 = 21·8$

and so on.

First moving average $= \frac{22·8}{5} = 4·56$

Second moving average $= \frac{21·8}{5} = 4·36$

and so on.]

EXAMPLE :

Calculate the four-yearly moving averages of the following :

Year :	1951	1952	1953	1954	1955	1956	1957	1958
Y :	506	620	1036	673	588	696	1116	738
		1959	1960	1961	1962	1963	1964	1965
		663	773	1189	818	745	845	1276

SOLUTION :

Caculation of 4-yearly Moving Averages

Year (1)	Y (2)	4-year moving total (3)	4-year moving average (4)	2-item moving total of col. (4) (5)	4-year centred moving average (6)
1951	506				
1952	620				
		2835	708·75		
1953	1036			1438·00	719·00
		2917	729·25		
1954	673			1477·50	738·75
		2993	748·25		
1955	588			1516·50	758·25
		3073	768·25		
1956	696			1552·75	776·38
		3138	784·50		
1957	1116			1587·75	793·87
		3213	803·25		
1958	738			1625·75	812·87
		3290	822·50		
1959	663			1663·25	831·67
		3363	840·75		
1960	773			1701·50	850·75
		3443	860·75		
1961	1189			1742·00	871·00
		3525	881·25		
1962	818			1780·50	890·25
		3597	899·25		
1963	745			1820·25	910·12
		3684	921·00		
1964	845				
1965	1276				

$$\text{col. (4)} = \frac{\text{col (3)}}{4}, \quad \text{col. (6)} = \frac{\text{col. (5)}}{2}$$

[**Working Notes.**

First 4-year moving total = 506 + 620 + 1036 + 673 = 2835

Second ... ... = 620 + 1036 + 673 + 588 = 2917 and so on.

First 2-item moving total = 708·75 + 729·25 = 1438·00

Second ... ... = 729·25 + 748·25 = 1477·50 and so on.]

[**Alternative Method**]

Calculation of 4-yearly Moving Averages

Year	Y	*4-year moving total*	*2-item moving total of col.* (3)	*4-year centred moving average* = $\frac{\text{col. (4)}}{8}$
(1)	(2)	(3)	(4)	(5)
1951	506			
1952	620	2835		
1953	1036	2917	5752	719·00
1954	673	2993	5910	738·75
1955	588	3073	6066	758·25
1956	696	3138	6211	776·38
1957	1116	3213	6351	793·87
1958	738	3290	6503	812·87
1959	663	3363	6653	831·67
1960	773	3443	6806	850·75
1961	1189	3525	6968	871·00
1962	818	3597	7122	890·25
1963	745	3684	7281	910·12
1964	845			
1965	1276			

Example :

From the following series of observations, calculate 5-yearly *weighted moving average* with weights 1, 2, 2, 2, 1 respectively.

Year :	1969	1970	1971	1972	1973	1974	1975	1976	1977
Annual sales (Rs. '0000) :	2	6	1	5	3	7	2	6	4

Year :	1978	1979
Annual sales (Rs. '0000) :	8	3

SOLUTION :

Compilation of 5-yearly Weighted Moving Average

Year (1)	*Annual sales* (Rs. '0000) (2)	*5-year weighted moving total* (3)	*5-year weighted moving average* (4)
1969	2		
1970	6		
1971	1	1.2 + 2.6 + 2.1 + 2.5 + 1.3 = 29	29 ÷ 8 = 3·625
1972	5	1.6 + 2.1 + 2.5 + 2.3 + 1.7 = 31	31 ÷ 8 = 3·875
1973	3	1.1 + 2.5 + 2.3 + 2.7 + 1.2 = 33	33 ÷ 8 = 4·125
1974	7	1.5 + 2.3 + 2.7 + 2.2 + 1.6 = 35	35 ÷ 8 = 4·375
1975	2	1.3 + 2.7 + 2.2 + 2.6 + 1.4 = 37	37 ÷ 8 = 4·625
1976	6	1.7 + 2.2 + 2.6 + 2.4 + 1.8 = 39	39 ÷ 8 = 4·875
1977	4	1.2 + 2.6 + 2.4 + 2.8 + 1.3 = 41	41 ÷ 8 = 5·125
1978	8		
1979	3		

Total weights = 1 + 2 + 2 + 2 + 1 = 8

(4) *Method of Least Square*

The method of least square is the most objective and widely used method in determining the trend in a time series data. When the data are plotted on the graph paper, it will be seen that all the points will not lie on a curve and quite a large number of curves can be drawn, by inspection, between the points. In order to find the best fitting curve to the data the method of least square is followed. This method consists in finding the best fitting curve to the time series data as that curve, from all possible curves, for which (i) the sum of the vertical deviations of the actual (observed) values from the fitted curve is zero and (ii) the sum of the squared vertical deviations is minimum, that is, no other curve would have a smaller sum of squared deviations. A graphical representation of the data is required to enable a decision to be made as to the particular curve to be fitted.

I. LINEAR TREND

When the trend is linear the trend equation may be represented by $y = a + bt$ and the values of a and b for the line $y = a + bt$ which minimizes the sum of squares of the vertical deviations of the actual (observed) values from the straight line, are the solutions to the so-called *normal equations* :

$$\Sigma y = na + b\,\Sigma t \quad \cdots \quad (1)$$

$$\Sigma yt = a\Sigma t + b\,\Sigma t^2 \quad \cdots \quad (2)$$

where n is the number of paired observations.

The normal equations are obtained by multiplying $y = a + bt$, by the coefficients of a and b, *i.e.*, by 1 and t throughout and summing up.

Case I. When the number of years is odd.

When the number of years is odd the calculation will be simplified by taking the mid-year as origin and one year as unit and in that case $\Sigma t = 0$ and the two normal equations take the form

$$\Sigma y = na$$

$$\Sigma yt = b\Sigma t^2$$

and hence, $$a = \frac{\Sigma y}{n},\ b = \frac{\Sigma yt}{\Sigma t^2}.$$

Case II. When the number of years is even.

When the number of years is even the origin is placed in the midway between the two middle years and the unit is taken to be $\frac{1}{2}$ year instead of one year With this change of origin and scale we have again $\Sigma t = 0$ and hence $a = \frac{\Sigma y}{n}$ and $b = \frac{\Sigma yt}{\Sigma t^2}$.

EXAMPLE :

Fit a straight line trend to the following data :

Year	: 1965	1966	1967	1968	1969	1970	1971
Gross ex-factory value (Rs. crores) :	672	824	967	1204	1464	1758	2057

and estimate the Gross ex-factory value (Rs. crores) for the year 1975.

[I. C. W. A. Dec. 1975]

SOLUTION :

Let the straight line trend is represented by the equation $y = a + bt$. The values of a and b shall be determined by solving the normal equations $\Sigma y = na + b\Sigma t$ and $\Sigma yt = a\Sigma t + b\Sigma t^2$.

Here, since the number of years is odd the mid-year, *i.e.*, year 1968 is taken as origin and one year as unit.

Fitting of Straight Line Trend

Year	Gross ex-factory value (Rs. crores) (y)	t = Year—1968	t^2	yt
1965	672	−3	9	−2016
1966	824	−2	4	−1648
1967	967	−1	1	−967
1968	1204	0	0	0
1969	1464	1	1	1464
1970	1758	2	4	3516
1971	2057	3	9	6171

$\therefore \quad \Sigma t=0,\ \Sigma t^2=28,\ \Sigma yt=6520,\ n=7,\ \Sigma y=8946$

From normal equations,

$8946=7a+b\times 0$ or, $8946=7a$ or, $a=1278$

$6520=a\times 0+b\times 28$ or, $6520=28b$ or, $b=232{\cdot}9$

$\therefore$ the trend equation is

$y=1278+232{\cdot}9\,t$ with origin at 1968 and t unit 1 year.

The value of t for 1975 will be 7. Hence the estimate for the year 1975 is

$$y=1278+232{\cdot}9\times 7=1278+1630{\cdot}3=2908{\cdot}3 \text{ (Rs. crores).}$$

EXAMPLE :

Fit a straight line trend to the following data :

Year	1951	1952	1953	1954	1955	1956
Electricity generated (million kw hours)	101	107	113	121	136	148

[I. C. W. A., Jan. 1967]

SOLUTION :

Let the straight line trend be represented by $y=a+bt$. Here the number of years is even, so we take the *origin at the mid-point* of 1953 and 1954 and t unit $=\frac{1}{2}$ year.

Fitting of Straight Line Trend

Year	Electricity generated (y)	$t=\dfrac{\text{Year}-1953{\cdot}5}{2}$	t^2	yt
1951	101	-5	25	-505
1952	107	-3	9	-321
1953	113	-1	1	-113
1954	121	1	1	121
1955	136	3	9	408
1956	148	5	25	740
Total	726	0	70	330

Normal equations :

$\Sigma y = an + b\Sigma t$ or, $726 = 6a + b\times 0$ or, $726 = 6a$ or, $a = 121$

$\Sigma yt = a\Sigma t + b\Sigma t^2$ or, $330 = a\times 0 + b\times 70$ or, $330 = 70b$ or, $b = 4{\cdot}71$

$\therefore$ the trend equation is

$y = 121 + 4{\cdot}71t$ with origin at the mid-point of 1953 and 1954 and t unit $= \frac{1}{2}$ year.

II. Fitting a Parabolic Trend

For a parabolic trend the equation is of the form

$$y = a + bt + ct^2$$

The normal equations to find the constants a, b and c are given by

$$\Sigma y = na + b\Sigma t + c\Sigma t^2$$
$$\Sigma yt = a\Sigma t + b\Sigma t^2 + c\Sigma t^3$$
$$\Sigma yt^2 = a\Sigma t^2 + b\Sigma t^3 + c\Sigma t^4$$

Taking the mid-point of the period as origin and unit as one year when n is odd and $\frac{1}{2}$ year when n is even, we have $\Sigma t = 0$ and $\Sigma t^3 = 0$ and the normal equations reduces to

$$\Sigma y = na + c\Sigma t^2$$
$$\Sigma yt = b\Sigma t^2$$
$$\Sigma yt^2 = a\Sigma t^2 + c\Sigma t^4$$

which can be solved easily for the three constants a, b and c.

EXAMPLE :

Fit by the method of least squares a parabolic curve to the following data :

Year :	1971	1972	1973	1974	1975	1976	1977
Production in tons :	23	20	18	18	14	13	13

SOLUTION :

Let the parabolic curve (trend) be represented by $y=a+bt+ct^2$. Since the number of years is odd the origin is taken at the mid-year, *i.e.*, 1974 and unit as one year.

Fitting of Parabolic Trend

Year	*Production* (y)	$t=$ *Year* —1974	t^2	t^3	t^4	yt	yt^2
1971	23	-3	9	-27	81	-69	207
1972	20	-2	4	-8	16	-40	80
1973	18	-1	1	-1	1	-18	18
1974	18	0	0	0	0	0	0
1975	14	1	1	1	1	14	14
1976	13	2	4	8	16	26	52
1977	13	3	9	27	81	39	117
Total	119	0	28	0	196	-48	488

From the normal equations

$$\Sigma y=an+b\Sigma t+c\Sigma t^2 \quad \cdots \quad \text{(i)}$$
$$\Sigma yt=a\Sigma t+b\Sigma t^2+c\Sigma t^3 \quad \cdots \quad \text{(ii)}$$
$$\Sigma yt^2=a\Sigma t^2+b\Sigma t^3+c\Sigma t^4 \quad \cdots \quad \text{(iii)}$$

Now, putting the values of $\Sigma t=0$, $\Sigma t^2=28$, $\Sigma t^3=0$, $\Sigma t^4=196$, $\Sigma yt=-48$ and $\Sigma yt^2=488$, we get,

$$119=7a+28c \quad \cdots \quad \text{(iv)}$$
$$-48=28b \quad \cdots \quad \text{(v)}$$
$$488=28a+196c \quad \cdots \quad \text{(vi)}$$

Solving the three equations,

$$a = 16{\cdot}43,$$
$$b = -1{\cdot}71,$$
$$c = 0{\cdot}143.$$

Thus,

$$y = 16{\cdot}43 - 1{\cdot}71\,t + 0{\cdot}143\,t^2$$

with origin 1974 and t unit one year.

III. Other Types of Trend Curves

(i) The following curve is often found to be a good fit to the time series data which is reducible to the linear form

$$y = a.b^t \text{ (exponential curve)}$$

which is reducible to the linear form by taking logarithm on both sides.

Thus, $\log y = \log a + t.\log b$.

A straight line trend to *log y* and t may be fitted and obtain the least square estimates of *log a* and *log b* and hence a and b.

(ii) In special cases more complicated curves are found to be a good fit to the time series data.

They are

(i) $y = \dfrac{1}{a + b.c^t}$ (Logistic curve)

(ii) $y = a.b^{c^t}$ (Gompertz curve)

(iii) $y = a + b.c^t$ (modified exponential curve).

Merits and Demerits.

(i) The mathematical curves fitted to the data is most suitable for forecasting purposes.

(ii) Being defined by a mathematical equation the method is most objective. There is no scope of any subjectiveness in this method.

(iii) This method involves more calculations as compared to other methods.

(iv) It is not flexible in the sense that an addition of some more values for some corresponding additional years changes the entire calculations of the trend equation.

Measurement of Seasonal Variations.

Methods used in measuring Seasonal Variations are as follows :

(1) *Method of Averages.*

(2) *Ratio to Trend Method.*

(3) *Moving Average Method.*

(4) *Link Relative Method.*

(1) *Method of Averages*

This method, the most simplest one, ignores any trend or any cyclical fluctuations that may be present in the series. It is applied only when the series does not contain trend or cyclical fluctuations to any appreciable extent and has stable, unchanging seasonal fluctuations. To find the seasonal variations, arrange the data by years and months, if monthly data are given, and obtain the total of the values for each month. Calculate the average value for all Januaries, all Februaries, etc. and then *average* these *averages*. If multiplicative model is used, then the percentage of each monthly average to average of monthly averages will be the seasonal indices. Sometimes slight adjustment is necessary to make the total indices 1200, as in this case 100 percent is considered to be the seasonal value for a normal month.

Instead of monthly figures if quarterly, weekly, etc. figures are given the same procedure explained above will be followed.

Since most of the time series data contains trend and cyclical fluctuations, this method is rarely used.

When additive model is used, the deviations of the *average* of monthly *averages* from each monthly averages will be the seasonal variations. In this case also, sometimes slight adjustments are to be made to make the total seasonal variations zero.

EXAMPLE :

Quarterly sales in (Rs. 000) of a company is given below :

Quarters	*Years*		
	1976	1977	1978
I	7·2	7·4	8·4
II	5·0	6·8	6·0
III	7·8	7·4	6·2
IV	9·2	9·0	7·6

Calculate the seasonal indices.

SOLUTION :

As no appreciable trend is noticed in the given data method of quarterly averages to be used here.

Calculation for Seasonal Indices

Quarter	*Year*			*Total*	*Average*
	1976	1977	1978		
I	7·2	7·4	8·4	23·0	7·67
II	5·0	6·8	6·0	17·8	5·93
III	7·8	7·4	6·2	21·4	7·13
IV	9·2	9·0	7·6	25·8	8·60
Total					29·33

$$\text{Average of averages} = \frac{29{\cdot}33}{4} = 7{\cdot}33$$

$$\text{Seasonal Index for Quarter I} = \frac{7{\cdot}67}{7{\cdot}33} \times 100 = 104{\cdot}6$$

$$\text{,, ,, ,, ,, II} = \frac{5{\cdot}93}{7{\cdot}33} \times 100 = 80{\cdot}9$$

$$\text{,, ,, ,, ,, III} = \frac{7{\cdot}13}{7{\cdot}33} \times 100 = 97{\cdot}2$$

$$\text{,, ,, ,, ,, IV} = \frac{8{\cdot}60}{7{\cdot}33} \times 100 = 117{\cdot}3$$

Note. Since the total of Seasonal Index for the 4 quarters is 400, no adjustment is necessary.

EXAMPLE :

Compute the average seasonal movements by method of average for the following data :

Year	*Quarters*			
	I	II	III	IV
1970	30	31	30	33
1971	34	27	18	24
1972	29	30	28	34
1973	31	29	25	30

SOLUTION :

Calculation for Average Seasonal Movement

Year	*Quarters*				*Total*
	I	II	III	IV	
1970	30	31	30	33	—
1971	34	27	18	24	—
1972	29	30	28	34	—
1973	31	29	25	30	—
Total	124	117	101	121	463
Average	31·0	29·25	25·25	30·25	115·75
Average Seasonal Movements	2·06	0·31	−3·69	1·32	0

$$\text{Average of averages of four quarters} = \frac{115{\cdot}75}{4} = 28{\cdot}94.$$

[Working Notes :

Average seasonal movements

for quarter I $= 31{\cdot}0 - 28{\cdot}94 = 2{\cdot}06$

for quarter II $= 29{\cdot}25 - 28{\cdot}94 = 0{\cdot}31$

for quarter III $= 25{\cdot}25 - 28{\cdot}94 = -3{\cdot}69$

for quarter IV $= 30{\cdot}25 - 28{\cdot}94 = 1{\cdot}31$]

Note : 1·31 has been arbitrarily changed to 1·32 to make the sum of the seasonal variations 0.

(2) *Ratio to Trend Method.*

Since most of the economic time series data contains trend to an appreciable extent, the method of averages is to be used after eliminating the trend from the time series data.

This method involves estimation of the trend, elimination of the trend from the given series and then to use the method of average to obtain the seasonal indices. This is done as follows :

Estimate the trend by fitting a mathematical curve and eliminate the trend from the given series by expressing the original data as percentage of the corresponding trend values. Arrange these percentages by years and months if monthly figures are given and obtain the average for each month and then *average* these twelve monthly averages. The percentages of each monthly average to the *average of monthly averages* will be the seasonal indices.

Same procedure to be followed when quarterly or weekly, etc. figures are available.

(3) *Moving Average Method.*

This method consists in eliminating the trend and cyclical components of the time series data and then to use the method of average to obtain an estimate of seasonal indices.

APPROACH FOR THIS METHOD :

Eliminate seasonal variations by taking a centred moving average with a period of 12 months if monthly data are involved. This moving average will also largely smooth out irregular movements. If, now, the multiplicative model is used then ratios to moving averages expressed as percentages, that is, original observations expressed as percentages of the corresponding 12 month moving averages will contain only the seasonal variations and irregular variations which has been eliminated by the moving average process. These percentages are then arranged by months and the average for each month is calculated. The seasonal indices are then found by expressing these monthly averages as percentage of the *average* of the 12 *monthly averages*, so that the sum of these indices is 1200.

Note. 1. If additive model is used then deviations in place of ratios are to be taken.

Note 2. If quarterly, weekly, etc. figures are given then also the same procedure will be followed.

EXAMPLE :

Using additive model compute seasonal variations of the following by the method of moving average and obtain deseasonalised data for the four quarters of 1973.

Quarterly Output of Paper in million tons

Years	I	II	III	IV
1971	37	38	37	40
1972	41	34	25	31
1973	35	37	35	41

SOLUTION :

Since quarterly figures are given a centred moving average with a pericd of four quarters is necessary.

Calculation of Moving Averages and Deviations

Years and Quarters		Quarterly output (million tons) (1)	4-quarterly moving total (2)	2-item centred moving total (3)	4-quarterly centred moving average $(4)=\left\{\frac{3}{8}\right\}$	Deviations (million tons) (5)=(1)−(4)
1971	I	37				
	II	38	152			
	III	37	156	308	38·5	−1·5
	IV	40	152	308	38·5	1·5
1972	I	41	140	292	36·5	4·5
	II	34	131	271	33·9	0·1
	III	25	125	256	32·0	−7·0
	IV	31	128	253	31·6	−0·6
1973	I	35	138	266	33·2	1·8
	II	37	148	286	35·7	1·3
	III	35				
	IV	41				

Calculation of Seasonal Variations

Years/Quarters	Deviations (million tons)				Total
	I	II	III	IV	
1971	—	—	−1·5	1·5	—
1972	4·5	0·1	−7·0	−0·6	—
1973	1·8	1·3	—	—	—
Total	6·3	1·4	−8·5	0·9	—
Average	3·15	0·7	−4·25	0·45	0·05
Adjustment	−0·0125	−0·0125	−0·0125	−0·0125	−0·05
Seasonal Variations	3·1375	0·6875	−4·2625	0·4375	0

$$(\text{Adjustment} = -\frac{0·05}{4} = -0·0125)$$

Deseasonalised data for 1973 :

For first quarter	35 − 3·1375 = 31·8625
For second quarter	37 − 0·6875 = 36·3125
For third quarter	35 + 4·2625 = 39·2625
For fourth quarter	41 − 0·4375 = 40·5625

EXAMPLE :

Compute Seasonal Indices of the following by the use of ratio-to-moving-average method.

Production of Paper in (thousand tons)

Years/Quarters	I	II	III	IV
1951	120	118	122	124
1952	124	122	126	129
1953	128	125	131	134
1954	132	129	135	138

SOLUTION :

Years and Quarters		*Data*	*4-quarter moving total*	*4-quarter moving average*	*Moving average of two moving averages*	*Ratio to moving average × 100*
1951	I	120				
	II	118				
			484	121·00		
	III	122			121·50	100·4
			488	122·00		
	IV	124			122·50	101·2
			492	123·00		
1952	I	124			123·50	100·4
			496	124·00		
	II	122			124·68	97·9
			501	125·25		
	III	126			125·75	100·1
			505	126·25		
	IV	129			126·62	101·8
			508	127·00		
1953	I	128			127·62	100·2
			513	128·25		
	II	125			128·88	97·5
			518	129·50		
	III	131			130·00	100·8
			522	130·50		
	IV	134			131·00	102·3
			526	131·50		
1954	I	132			132·00	100·0
			530	132·50		
	II	129			133·00	96·9
			534	133·50		
	III	135				
	IV	138				

Calculation of Seasonal Indices

Years/Quarters	*Ratio to moving average* × 100				*Total*
	I	II	III	IV	
1951	—	—	100·4	101·2	—
1952	100·4	97·9	100·1	101·8	—
1953	100·2	97·5	100·8	102·3	—
1954	100·0	- 96·9	—	—	—
Total	300·6	292·3	301·3	305·3	—
A.M.	100·2	97·4	100·4	101·8	399·8
Seasonal Indices	100·25	97·44	100·45	101·86	400·0

[**Working Notes :**

Average of A.M.'s $=\frac{399\cdot8}{4}=99\cdot95.$

Seasonal Indices :

For first quarter $=\frac{100\cdot2}{99\cdot95}\times100=100\cdot25$

For second quarter $=\frac{97\cdot4}{99\cdot95}\times1000=97\cdot44$

and so on.

The sum of Seasonal Indices for four quarters must be 400]

Conversion of Annual Trend to Monthly Values :

In time series annual data are usually employed to compute the trend. It is sometime necessary to obtain monthly trend values instead of yearly values from the trend line.

The annual data employed may refer to annual totals or may refer to monthly averages for each year obtained from annual totals by dividing each total for a year by 12.

(i) *When Annual Data are Annual Totals :*

Let $y=a+bt$ be the least square trend fitted to the annual totals for some years.

Taking the mid-year as origin and one year as unit when n, the number of years covered, is *odd* we have $a=\frac{\Sigma y}{n}$ which is the arithmetic mean of n yearly totals. This value of a for annual totals when taken in monthly terms would be $\frac{1}{12}$th of it. From annual data the term

b is the trend increment for entire year. Dividing this **b** by 12 we have monthly trend increment in the yearly totals. Since still we have yearly totals **b** should be divided again by 12 to reduce it in monthly terms.

Thus the monthly equation would be

$$y=\frac{a}{12}+\frac{b}{144}\,t=c+dt$$

(origin : mid-year, *i.e.*, June-July ; t unit = 1 month)

Since the number of months in a year is even, the origin is in the middle of two months. Thus **c** is the value of trend at the end of June. Since monthly trend values should refer to the middle of the month, the origin should be shifted from the middle of the two months to the mid-point of any convenient month. If middle of July is taken as origin the trend equation for monthly values will be

$$y=c+d\,(t+\tfrac{1}{2})$$

$$=\frac{a}{12}+\frac{b}{144}\,(t+\tfrac{1}{2})$$

(origin : July ; t unit = 1 month)

by shifting the origin half month later.

Again when the number of years covered is *even*, as explained earlier the t unit will be of 6 months and hence the monthly trend increment here will be $\frac{\mathbf{b}}{6}$ as **b** is the trend increment for 6 months. Hence proceeding in the same way as just described except for the fact that instead of dividing **b** by 144 it is to be divided now by 6 × 12 = 72. Hence the monthly trend equation is

$$y=\frac{a}{12}+\frac{b}{72}\,t$$

(origin : Dec.-January ; t unit = 1 month)

Now if January is taken as origin then monthly trend equation becomes

$$y=\frac{a}{12}+\frac{b}{72}\,(t+\tfrac{1}{2})$$

(origin : January ; t unit = 1 month)

by shifting the origin half month later.

(ii) *When Annual Data are Monthly Averages :*

When the trend has been fitted to the annual data which are monthly averages, it is simply required to divide **b**, the annual trend increment, by 12 when the number of years covered is *odd* and to divide **b** by 6 when the number of years covered is *even* and then shift the origin to the middle of any convenient month in the same way as described in (i).

Thus the monthly trend equation when the number of years covered is *odd*, is

$$y=a+\frac{b}{12}\,(t+\tfrac{1}{2})$$

(origin : July ; t unit = 1 month)

and monthly trend equation when the number of years covered is *even* is,

$$y=a+\frac{b}{6}\,(t+\tfrac{1}{2})$$

(origin : January ; t unit = 1 month)

EXAMPLE :

Find the Seasonal Indices from the following data by the ratio to trend method :

Years	Qtr. I	Qtr. II	Qtr. III	Qtr. IV
1971	45	60	54	51
1972	51	78	75	66
1973	60	87	81	72
1974	81	114	102	93
1975	120	138	129	123

SOLUTION :

Calculation of Trend by Method of Least Square

Years	Yearly totals	Quarterly average (y)	t = Years − 1973	yt	t^2	Trend Values
1971	210	52·5	−2	−105·0	4	48
1972	270	67·5	−1	−67·5	1	66
1973	300	75·0	0	0	0	84
1974	390	97·5	1	97·5	1	102
1975	510	127·5	2	255·0	4	120

$\Sigma y = 420{\cdot}0,\quad \Sigma yt = 180{\cdot}00,\quad \Sigma t^2 = 10,\quad n=5,\quad \Sigma t = 0$

Let $y=a+bt$ be the equation of straight line trend, then the normal equations are given by,

$$\Sigma y = na + b\Sigma t$$
$$\Sigma yt = a\Sigma t + b\Sigma t^2$$

which gives,

$$\left.\begin{array}{l}420=5a+b\times 0\\180=a\times 0+10b\end{array}\right\}\text{ or, }\left.\begin{array}{r}5a=420\\10b=180\end{array}\right\}\text{ or }\begin{array}{l}a=84\\b=18\end{array}$$

The yearly trend increment $=18$

$\therefore$ quarterly trend increment $=\frac{18}{4}=4·5$.

Calculation of Quarterly Trend Values: The straight line trend has been fitted to the annual data which are quarterly averages. So, for the year, say 1971, the trend value for the middle quarter is 48, that is, the trend value for half of second quarter and half of third quarter is 48. Hence the trend value for the second quarter is $48-\frac{1}{2}\times 4·5=48-2·25=45·75$ and for third quarter is $48+\frac{1}{2}\times 4·5=50·25$. The trend value for first quarter is $45·75-4·5=41·25$ and for third quarter $50·25+4·5=54·75$, and so on.

	Trend Values			
Year	Qtr. I	Qtr. II	Qtr. III	Qtr. IV
1971	41·25	45·75	50·25	54·75
1972	59·25	63·75	68·25	72·75
1973	77·25	81·75	86·25	90·75
1974	95·25	99·75	104·25	108·75
1975	113·25	117·75	122·25	126·75

Calculation of Seasonal Indices

	Ratio to trend expressed as %				
Year	Qtr. I	Qtr. II	Qtr. III	Qtr. IV	Total
1951	109·1	131·2	107·6	93·2	—
1952	86·0	122·3	109·8	90·6	—
1953	77·6	106·4	93·9	79·3	—
1954	85·1	114·3	97·8	85·4	—
1955	106·1	117·2	105·5	97·1	—
Total	463·9	591·4	514·6	445·6	—
A.M.	92·78	118·28	102·92	89·12	403·10
Seasonal Indices	92·1	117·3	102·2	88·4	400·00

[Average of A.M.'s $= \frac{403\cdot10}{4} = 100\cdot775$

Seasonal Indices, for first quarter $= \frac{92\cdot78}{100\cdot775} \times 100 = 92\cdot1$

for second quarter $= \frac{118\cdot28}{100\cdot775} \times 100 = 117\cdot3$

and so on.]

(4) *Link Relative Method.*

This method is the most difficult one and based upon the averages of Link Relatives. Briefly, the method is as follows :

(1) Calculate the link relatives by expressing the value for a month, if monthly figures are given, as percentage of the immediately preceding month's value. The link relative for the first month of the first year can not be obtained.

(2) Arrange the link relatives by months and calculate the average link relatives for each month using preferably the median. The arithmetic mean can also be used to calculate the averages. Let these averages be A_1, A_2, ..., A_{12}.

(3) Convert these averages into *Chain Relatives* by relating them to a common base, that is the first month January. The chain relative for this January is taken as 100 per cent and the chain relative for any month is now obtained by multiplying link relatives of that month by chain relative of previous month and dividing the product by 100. Let the chain relatives be $c_1, c_2, ..., c_{12}, c_{13}$—where c_{13} is the second chain relative for January.

(4) There will be some difference between c_1 and c_{13} and that difference is due to the presence of *Trend*. It is, therefore, necessary to adjust for trend. For this, the difference $c_{13}-c_1$ is considered to be the annual trend increment. So, the monthly trend increment is $\frac{c_{13}-c_1}{12} = c$ (say). This monthly increment c multiplied by 1, 2, 3, ..., 11 is deducted from the chain relatives c_2, ..., c_{12} respectively to get adjusted (corrected) chain relatives.

(5) These adjusted (corrected) chain relatives are then expressed as percentages of their average to provide the required seasonal indices. The sum of the seasonal indices will be 1200.

Note. If quarterly, weekly, etc. figures are given the same procedure explained above to be followed.

Example :

Calculate seasonal indices by the method of link relatives from the following data :

Production in tons

Years	1st Qtr.	2nd Qtr.	3rd Qtr.	4th Qtr.
1976	360	364	388	380
1977	364	380	398	388
1978	366	412	436	414
1979	366	416	450	440

Solution :

Seasonal Indices by Method of Link Relatives

Years/Quarters	Link Relatives				
	I	II	III	IV	I
1976	—	101·1	106·6	97·9	—
1977	95·8	104·4	104·7	97·5	—
1978	94·3	112·6	105·8	94·9	—
1979	88·4	113·7	108·2	97·8	—
Total	278·5	431·8	425·3	388·1	—
A.M.	92·83	107·95	106·32	97·02	—
C.R.	100	107·95	114·77	111·35	103·36
Adjusted C.R.	100	107·11	113·09	108·83	100
Seasonal Indices	93·2	99·9	105·4	101·5	—

Steps :

(i) L.R : 364 ÷ 360 = 101·1, 388 ÷ 364 = 106·6, etc.

(ii) C.R : for 1st Qtr. $= 100$ (assumed) ;

$$\text{for 2nd Qtr.} = \frac{100 \times 107·95}{100} = 107·95\,;$$

$$\text{for 3rd Qtr.} = \frac{106·32 \times 107·95}{100} = 114·77\,;$$

$$\text{for 4th Qtr.} = \frac{114·77 \times 97·02}{100} = 111·35\,;$$

$$\text{for 1st Qtr. (Second C.R.)} = \frac{111·35 \times 92·83}{100} = 103·36.$$

(iii) Adjustment (correction) factor $= \dfrac{103·36 - 100}{4} = 0·84.$

Adjusted (corrected) C.R. : for 1st Qtr. $= 100$;
for 2nd Qtr. $= 107·95 - 0·84 = 107·11$;
for 3rd Qtr. $= 114·77 - 1·68 = 113·09$;
for 4th Qtr. $= 111·35 - 2·52 = 108·83$;
for 1st Qtr. (Second C.R.) $= 103·36 - 3·36 = 100.$

(iv) A.M. of adjusted C.R.'s $= \dfrac{100 + 107·11 + 113·09 + 108·83}{4}$
$= 107·26.$

Seasonal Indices : for 1st Qtr. $= \dfrac{100}{107·26} \times 100 = 93·2$;

for 2nd Qtr. $= \dfrac{107·11}{107·26} \times 100 = 99·9$;

for 3rd Qtr. $= \dfrac{113·09}{107·26} \times 100 = 105·4$;

for 4th Qtr. $= \dfrac{108·83}{107·26} \times 100 = 101·5.$

Total of seasonal indices should be 400.

Note. *The averaging process used in all the methods explained above is only to smooth out the irregular variations as far as possible.*

Determination of Cyclical Variations.

Residual Method : This method is most commonly used and consists in eliminating the trend, the seasonal variations and irregular variations from the time series data in order to isolate the Cyclical Variations. The usual procedure involves, first, in eliminating seasonal variations from the data giving the deseasonalised data and then extracting trend from the deseasonalised data, leaving only the cyclical-irregular variations. However, the trend may be removed first from the data, giving the trend adjusted data and, next, seasonal variations may be eliminated, leaving only the cyclical-irregular variations. Another possibility is to obtain the product, for multiplicative model (or sum, for additive model) of the trend and seasonal values and to eliminate both of these movements at the same time, from the time series data, leaving again only the cyclical-irregular variations.

The cyclical-irregular movements thus obtained may further be smoothed by the method of moving average in order to obtain cyclical variations. The irregular variations, in general, cannot be eliminated completely but can be smoothed, so as to bring a better picture of the cyclical variations by the use of short-term weighted moving averages.

Miscellaneous Examples

1. Calculate a five-year moving average of production data given below :

Year :	1929	1930	1931	1932	1933	1934	1935	1936	1937	1938
Production :	1590	1516	1364	1134	1063	1302	1428	1773	1917	1990

(C. U. 1966)

SOLUTION :

Compilation of Five-year Moving Average

Years	*Production*	*5-year moving total*	*5-year moving average*
1929	1590	—	—
1930	1516	—	—
1931	1364	6667	1333·4
1932	1134	6379	1275·8
1933	1063	6291	1258·2
1934	1302	6700	1340·0
1935	1428	7483	1496·7
1936	1773	8410	1682·0
1937	1917	—	—
1938	1990	—	—

2. The following data give daily sales of a shop observing a five-day week, over four successive weeks. Determine the period of the moving average and calculate the moving averages accordingly.

Day :	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Sales :	26	29	35	47	51	26	32	37	46	53	28	30	36	46	54	28	31	36	46	54

(C. A. 1974)

SOLUTION :

The data show a regular cycle of 5 days and hence the period of moving average should be of 5-day.

Calculations for 5-day Moving Average

Day	*Sales*	*5-day moving total*	*5-day moving average*
1	26	—	—
2	29	—	—
3	35	188	37·6
4	47	188	37·6
5	51	191	38·2
6	26	193	38·6
7	32	192	38·4
8	37	194	38·8
9	46	196	39·2
10	53	194	38·8
11	28	193	38·6
12	30	193	38·6
13	36	194	38·8
14	46	194	38·8
15	54	195	39·0
16	28	195	39·0
17	31	195	39·0
18	36	195	39·0
19	46	—	—
20	54	—	—

3. Assuming a four-yearly cycle, calculate the trend by the method of moving average from the following data :

Year	:	1941	1942	1943	1944	1945	1946	1947	1948	1949	1950
Production	:	464	515	518	467	502	540	557	571	586	612

(I. C. W. A., 1968)

SOLUTION :

Here the cycle is of a period of 4 years. So, to obtain the trend values by moving average method, the period of moving average must be 4 years.

As the period of moving average is even, it is necessary to *centre* the moving averages.

Calculation of Trend Values by Moving Average

Year (1)	Production (2)	4-year moving total (3)	2-item moving total (4)	4-year moving average (centred) Trend Values (5)=(4)÷8
1941	464			
1942	515	1964		
1943	518	2002	3966	495·8
1944	467	2027	4029	503·6
1945	502	2066	4093	511·6
1946	540	2170	4236	529·5
1947	557	2254	4424	553·0
1948	571	2326	4580	572·5
1949	586			
1950	612			

4. Given below the production of coal in thousands of tons for the years 1971—75

Year	:	1971	1972	1973	1974	1975
Production	:	44·5	38·9	38·1	32·6	38·7

Use the method of least squares to fit a line to the data given above. What is the trend value in the year 1973 ?

[I. C. W. A., 1979]

SOLUTION :

Taking the origin at the year 1973 and unit of time $t=1$ year, let $y=a+bt$ be the equation of the trend line, where y denotes the production in thousands of tons. The constants a and b are to be determined by solving the normal equations, given by,

$$\Sigma y=a.n+b\Sigma t\ ;\ \Sigma yt=a\Sigma t+b\Sigma t^2 \qquad \cdots \quad (1)$$

Calculations for Linear Trend

Year	t	y	t^2	yt
1971	−2	44·5	4	−89·0
1972	−1	38·9	1	−38·9
1973	0	38·1	0	0
1974	1	32·6	1	32·6
1975	2	38·7	4	77·4

$\Sigma t=0$; $\Sigma y=192{\cdot}8$; $\Sigma t^2=10$; $\Sigma yt=-17{\cdot}9$; $n=5$.

Solving the normal equations :

$\Sigma y=an+b\Sigma t$ $\qquad$ $\Sigma yt=a\Sigma t+b\Sigma t^2$

$192{\cdot}8=a.5+b.0$ $\qquad$ $-17{\cdot}9=a.0+b.10$

or, $192{\cdot}8=5a$ $\qquad$ or, $-17{\cdot}9=10b$

or, $a=38{\cdot}56$ $\qquad$ or, $b=-1{\cdot}79.$

∴ the equation for trend is $y=38{\cdot}56-1{\cdot}79t$ with origin at 1973 and t units $=1$ year.

To find the trend value in the year 1973, put $t=0$ in the above equation, *i.e.*, $y_{1973}=38{\cdot}56-1{\cdot}79\times 0=38{\cdot}56$ thousand tons.

5. The weights (in lbs.) of a new-born calf are taken at weekly intervals. Below are the observations for 10 weeks :

Age (x) :	1	2	3	4	5	6	7	8	9	10
Weight (y) :	52·5	58·7	65·0	70·2	75·4	81·1	87·2	95·5	102·2	108·4

Let $y=a+bu$, where $u=2x-11$. Use normal equations to estimate a and b. (*Given* : The sum of the products of the form uy for these 10 observations $=1016{\cdot}8$). Hence obtain the line of best fit of y on x. Now, write down the average rate of growth of the calf per week.
[I. C. W. A., 1979]

SOLUTION :

The normal equations for determining the constants a and b are

$$\Sigma y = na + b\Sigma u \qquad \cdots \quad \text{(i)}$$
$$\Sigma yu = a\Sigma u + b\Sigma u^2 \qquad \cdots \quad \text{(ii)}$$

Calculation for Trend Equation

x	y	$u = 2x - 11$	u^2
1	52·5	−9	81
2	58·7	−7	49
3	65·0	−5	25
4	70·2	−3	9
5	75·4	−1	1
6	81·1	1	1
7	87·2	3	9
8	95·5	5	25
9	102·2	7	49
10	108·4	9	81
Total	796·2	0	330

Number of observations $= n = 10$. Also, given $\Sigma uy = 1016{\cdot}8$

Putting the values from the table in the normal equations,

$$\left.\begin{aligned} 796{\cdot}2 &= 10a + b.0 \\ 1016{\cdot}8 &= a.0 + 330b \end{aligned}\right\} \text{ or } \quad \begin{aligned} 10a &= 796{\cdot}2 \\ 330b &= 1016{\cdot}8 \end{aligned}$$

Solving, $a = 79{\cdot}62$; $b = 3{\cdot}08$.

∴ the line of best fit is then,

$$\begin{aligned} y &= 79{\cdot}62 + 3{\cdot}08u \\ &= 79{\cdot}62 + 3{\cdot}08\,(2x - 11) \\ &= 45{\cdot}74 + 6{\cdot}16x. \end{aligned}$$

The average rate of growth of the calf per week is 6·16 lbs.

6. Fit a straight line trend by the method of least squares and estimate the trend values :

Year :	1961	1962	1963	1964	1965	1966	1967	1968
Value :	80	90	92	83	94	99	92	104

[C. A., 1976]

SOLUTION :

Let $y=a+bt$ be the equation of straight line trend with origin at the mid-point of 1964 and 1965, and unit of time $t=\frac{1}{2}$ year.

The normal equations for finding the values of a and b are

$$\Sigma y=an+b\Sigma t$$
$$\Sigma yt=a\Sigma t+b\Sigma t^2.$$

Calculations for Straight Line Trend

Year	Value (y)	t	t^2	yt
1961	80	−7	49	−560
1962	90	−5	25	−450
1963	92	−3	9	−276
1964	83	−1	1	−83
1965	94	1	1	94
1966	99	3	9	297
1967	92	5	25	460
1968	104	7	49	728
Total	734	0	168	210

Putting the values from the table in the normal equations and noting that $n=8$,

$734=8a+b.0$; $\quad$ $210=a.0+168b.$

or, $8a=734$; $\quad$ or, $168b=210$

or, $a=91{\cdot}75$; $\quad$ or, $b=1{\cdot}25$

The trend equation is then,

$$y=91{\cdot}75+1{\cdot}25t$$

with origin at the mid-point of 1964 and 1965 and t units $=\frac{1}{2}$ year.

Calculation of Trend Values

Trend values for 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968 are obtained by substituting the corresponding values of t, viz., −7, −5, −3, −1, 1, 3, 5, 7, in the trend equation.

Trend value for 1961 $=91{\cdot}75+1{\cdot}25\times(-7)=83{\cdot}0$

1962 $=91{\cdot}75+1{\cdot}25\times(-5)=85{\cdot}5$

1963 $=91{\cdot}75+1{\cdot}25\times(-3)=88{\cdot}0$

1964 $=91{\cdot}75+1{\cdot}25\times(-1)=90{\cdot}5$

1965 $=91{\cdot}75+1{\cdot}25\times1 \quad =93{\cdot}0$

$$1966 = 91{\cdot}75 + 1{\cdot}25 \times 3 \quad = 95{\cdot}5$$
$$1967 = 91{\cdot}75 + 1{\cdot}25 \times 5 \quad = 98{\cdot}0$$
$$1968 = 91{\cdot}75 + 1{\cdot}25 \times 7 \quad = 100{\cdot}5.$$

7. Calculate the Seasonal Indices from the following data using the average method :

Years	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1974	72	68	80	70
1975	76	70	82	74
1976	74	66	84	80
1977	76	74	84	78
1978	78	74	86	82

(C. A., 1979)

SOLUTION :

Calculation of Seasonal Indices

Years	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1974	72	68	80	70
1975	76	70	82	74
1976	74	66	84	80
1977	76	74	84	78
1978	78	74	86	82
Total	376	352	416	384

$$\text{Quarterly average : 1st Quarter} = \frac{376}{5} = 75{\cdot}2\,;$$

$$\text{2nd Quarter} = \frac{352}{5} = 70{\cdot}4\,;$$

$$\text{3rd Quarter} = \frac{416}{5} = 83{\cdot}2\,;$$

$$\text{4th Quarter} = \frac{384}{5} = 76{\cdot}8.$$

$$\text{Average for all the data} = \frac{75{\cdot}2 + 70{\cdot}4 + 83{\cdot}2 + 76{\cdot}8}{4} = 76{\cdot}4.$$

$$\text{Seasonal Indices}: \text{ 1st Quarter} = \frac{75{\cdot}2}{76{\cdot}4} \times 100 = 98{\cdot}43\ ;$$

$$\text{2nd Quarter} = \frac{70{\cdot}4}{76{\cdot}4} \times 100 = 92{\cdot}14\ ;$$

$$\text{3rd Quarter} = \frac{83{\cdot}2}{76.4} \times 100 = 108{\cdot}90\ ;$$

$$\text{4th Quarter} = \frac{76{\cdot}8}{76{\cdot}4} \times 100 = 100{\cdot}53.$$

Note. The total of Seasonal Indices for the 4 quarters being 400, no adjustment is necessary.

8. A large company estimates its monthly average sales in a particular year to be Rs. 2,00,000. The Seasonal Indices of the sales data are as follows :

Month	Jan.	Feb.	March	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Seasonal Indices	76	77	98	128	137	122	101	104	100	102	82	73

Using the information, draw up a monthly sales budget for the company (assuming that there is no trend).

(C. A., 1978)

SOLUTION :

Average monthly sales = Rs. 2,00,000

$$\text{Estimated sales} = \frac{\text{Average Monthly Sales} \times \text{Monthly Seasonal Index}}{\text{Average Seasonal Index}}$$

$$\text{Average of the Seasonal Indices} = \tfrac{1}{12}(76+77+98+128+137+122 +101+104+100+102+82+73)$$

$$= \frac{1200}{12} = 100.$$

[Since seasonal indices are generally expressed in percentages, the average seasonal index is 100]

∴ Monthly Sales Budget :

Month	*Seasonal Indices*	*Estimated Sales* (Rs.)
January	76	(2,00,000 × 76) ÷ 100 = 1,52,000
February	77	(2,00,000 × 77) ÷ 100 = 1,54,000
March	98	(2,00,000 × 98) ÷ 100 = 1,96,000
April	128	(2,00,000 × 128) ÷ 100 = 2,56,000

Month	*Seasonal Indices*	*Estimated Sales* (Rs.)
May	137	(2,00,000 × 137) ÷ 100 = 2,74,000
June	122	(2,00,000 × 122) ÷ 100 = 2,44,000
July	101	(2,00,000 × 101) ÷ 100 = 2,02,000
August	104	(2,00,000 × 104) ÷ 100 = 2,08,000
September	100	(2,00,000 × 100) ÷ 100 = 2,00,000
October	102	(2,00,000 × 102) ÷ 100 = 2,04,000
November	82	(2,00,000 × 82) ÷ 100 = 1,64,000
December	73	(2,00,000 × 73) ÷ 100 = 1,46,000
		Total = 24,00,000

9. Using 4 quarterly moving average in respect of the following data, find (a) the trend, (b) short-term fluctuations, and (c) seasonal variations :

Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1971	35	86	67	124
1972	38	109	91	176
1973	47	158	104	226
1974	61	177	134	240
1975	72	206	141	307

(C. A. 1977)

SOLUTION :

Calculation of Moving Average (Trend) and Short-term fluctuations

Year/Quarter (1)		Data (2)	4-quarter moving total (3)	2-item moving total (4)	4-quarter moving average (Trend) (5) = (4) ÷ 8	Short-term fluctuations (6)
1971	1	35				
	2	86	312			
	3	67	315	627	78·37	−11·37
	4	124	338	653	81·63	+42·37
1972	1	38	362	700	87·50	−49·50
	2	109	414	776	97·00	+12·00
	3	91	423	837	104·63	−13·63
	4	176	472	895	111·88	+64·12
1973	1	47	485	957	119·63	−72·63
	2	158	535	1020	127·50	+30·50
	3	104	549	1084	135·50	−31·50
	4	226	568	1117	139·63	+86·37
1974	1	61	598	1166	145·75	−84·75
	2	177	612	1210	151·25	+25·75
	3	134	623	1235	154·38	−20·38
	4	240	652	1275	159·38	+80·62
1975	1	72	659	1311	163·88	−91·88
	2	206	726	1385	173·13	+32·87
	3	141				
	4	307				

Calculations for Seasonal Variations

Year	*1st Quarter*	*2nd Quarter*	*3rd Quarter*	*4th Quarter*
1971	—	—	−11·37	42·37
1972	−49·50	12·00	−13·63	64·12
1973	−72·63	30·50	−31·50	86·37
1974	−84·75	25·75	−20·38	80·62
1975	−91·88	32·87	—	—
Total	−298·76	101·12	−76·88	273·48
Average	−74·69	25·28	−19·22	68·37
Adjustment	·065	·065	·065	·065
Seasonal Variations	−74·625	25·345	−19·155	68·435

10. Fit a straight line trend to the following series of production data :

Year :	1960	1961	1962	1963	1964	1965	1966
Y	37	38	37	40	41	45	50

Y values being the monthly average production in thousand tons, what is the monthly trend increment ? Find the monthly trend values from the fitted equation for January and March of 1961. (C.U.)

SOLUTION :

Here the production is given in monthly average. Let the equation of straight line trend fitted to the yearly data be $y=a+bt$ with origin 1963 ; t unit = 1 year and y unit = thousand tons. The normal equations for estimating a and b are $\Sigma y=an+b\Sigma t$ and $\Sigma yt=a\Sigma t+b\Sigma t^2$.

Computation of Straight Line Trend

Year	y	t	t^2	yt
1960	37	−3	9	−111
1961	38	−2	4	−76
1962	37	−1	1	−37
1963	40	0	0	0
1964	41	1	1	41
1965	45	2	4	90
1966	50	3	9	150
Total	288	0	28	57

From normal equations,

$$\Sigma y=an+b\Sigma t \qquad \Sigma yt=a\Sigma t+b\Sigma t^2$$

$$288=7a+b.0 \qquad 57=a.0+28b$$

$$\text{or, } 288=7a \qquad \text{or, } 57=28b$$

$$\text{or, } a=41{\cdot}14 \qquad \text{or, } b=2{\cdot}04.$$

∴ the trend equation fitted to *yearly* data is

$y=41{\cdot}14+2{\cdot}04t$. (origin 1963 ; t unit = 1 year)

Since y represents the monthly average of production for each year and the unit of t is 1 year, *i.e.*, 12 months, so the trend of monthly average production increases by 2·04 in 12 months, *i.e.*, $2{\cdot}04\div12=0{\cdot}17$ per month. Hence, the monthly trend increment of production is 0·17 thousand tons.

The trend equation fitted to the *yearly* data is

$y=41{\cdot}14+2{\cdot}04\,t$ (origin : 1963 ; t unit = 1 year)

The trend equation fitted to the *monthly* data is

$y=41{\cdot}14+0{\cdot}17\,t$ (origin : June-July, 1963 ; t unit = 1 month)

For estimating monthly trend values, the origin must be shifted to the middle of a month. If July 1963 is to be taken as origin then origin to be shifted half a month later.

∴ the *monthly* trend equation fitted to the middle of July, 1963 is,

$y=41{\cdot}14+0{\cdot}17(t+\tfrac{1}{2})=41{\cdot}225+0{\cdot}17\,t$

(origin : July 1963, t unit = 1 month)

Estimation of Monthly Trend Values :

(i) January, 1961 is 30 months behind the origin and putting -30 for t in the trend equation for *monthly* values, we get,

$y=41{\cdot}225+0{\cdot}17(-30)=36{\cdot}125$ thousand tons.

(ii) March, 1961 is 28 months behind the origin and hence putting -28 for t in *monthly* trend equation, we have,

$y=41{\cdot}225+0{\cdot}17(-28)=36{\cdot}465$ thousand tons.

Ans. : (1) Monthly trend increment = 0·17 thousand tons.

(2) Trend value for January, 1961 = 36·125 thousand tons.

(3) Trend value for March, 1961 = 36·465 thousand tons.

11. Sales of a company rose from Rs. 39,45,000 to Rs. 46,21,000 from second quarter to third quarter. The seasonal indices for these quarters are 103 and 150 respectively. The owner of the company holds that it is a losing concern. Analyse the above information for supporting the owner's view. (C.U. 1967)

Solution :

Since the actual sales of the company for the second quarter were Rs. 39,45,000 and the seasonal index for that quarter is 103, the *normal* quarterly sales would be,

$$\text{Rs. } 39{,}45{,}000\times\frac{100}{103}=\text{Rs. } 38{,}30{,}097$$

and the *expected* sales for the third quarter would be,

$$\text{Rs. } 38{,}30{,}097\times\frac{150}{100}=\text{Rs. } 57{,}45{,}146$$

as the seasonal index for the third quarter is 150.

So the actual sales of the third quarter is far less than the expected sales of the same quarter.

Thus, the owner's view that the company is a losing concern is justified.

12. Fit an exponential trend $y=ab^t$ to the following data by the method of least squares and find the trend value for the year 1977.

Year (x) :	1971	1972	1973	1974	1975
Production (million tons) :	132	142	157	170	191

SOLUTION :

Since the number of years covered is odd, the *origin* is taken at the year 1973 and 1 *year* as unit. The exponential trend $y=ab^t$ is transformed into a *linear form* by taking logarithm on both sides. Thus, $\log y=\log a+t.\ \log b$, or, $Y=A+Bt$, where $Y=\log y$, $A=\log a$ and $B=\log b$. The constants A and B to be determined by solving the normal equations,

$$\Sigma Y=n.A+B\Sigma t, \quad \Sigma Yt=A\Sigma t+B\Sigma t^2.$$

Calculations for Fitting Exponential Trend

x	y	$Y=\log y$	$t=x-1973$	t^2	Yt
1971	132	2·1206	−2	4	−4·2412
1972	142	2·1523	−1	1	−2·1523
1973	157	2·1959	0	0	0
1974	170	2·2304	1	1	2·2304
1975	191	2·2810	2	4	4·5620

$\therefore\ \Sigma Y=10{\cdot}9802\ ;\quad \Sigma t=0\ ;\quad \Sigma t^2=10\ ;\quad \Sigma Yt=0{\cdot}3989\ ;\quad n=5.$

Substituting the values obtained from the table in the normal equations,

$$\begin{cases}10{\cdot}9802=5.A+\ 0.B\\ 0{\cdot}3989=0.A+10.B.\end{cases} \quad \text{or,} \quad \begin{cases}5A=10{\cdot}9802\\ 10B=\ 0{\cdot}3989.\end{cases}$$

Therefore, $A=2{\cdot}19604$, or, $\log a=2{\cdot}19604$, or, $a=\text{Antilog}\ (2{\cdot}19604)$
$=157{\cdot}01$

and $B=0{\cdot}03989$, or, $\log b=0{\cdot}03989$, or, $b=\text{Antilog}\ (0{\cdot}03989)$
$=1{\cdot}0962$

$\therefore$ the equation for exponential trend is $y=157{\cdot}01\ (1{\cdot}0962)^t$ with origin at 1973 and t units $=1$ year.

To find the trend value for the year 1977, put $t=4$ (since the value of t for the year 1977 is 4) in the equation $Y=A+Bt$.

$\therefore$ the trend value for the year 1977 is,

$$Y=2{\cdot}19604+0{\cdot}03989\times4=2{\cdot}35560$$

or, $\log y=2{\cdot}35560$, or, $y=\text{Antilog}\ (2{\cdot}35560)=226{\cdot}8$ million tons.

EXERCISE 12

1. During two consecutive weeks the attendances at an exhibition are recorded, the numbers being given 000's :

Week	*Sun.*	*Mon.*	*Tue.*	*Wed.*	*Thu.*	*Fri.*	*Sat.*
1	24	55	29	48	52	55	61
2	27	52	32	43	53	56	65

Calculate a seven-day moving average. (C. A. 1963)

[*Ans.* : 46·3, 46·7, 46·3, 46·7, 46·0, 46·1, 46·3, 46·9]

2. Compute 4-yearly moving averages from the following :

Year :	1930	1931	1932	1933	1934	1935	1936	1937
Value (Rs.) :	365	360	355	330	300	330	340	290
Year :	1938	1939	1940	1941	1942	1943	1944	1945
Value (Rs.) :	280	250	235	255	250	245	225	210
Year :	1946	1947	1948	1949	1950			
Value (Rs.) :	200	230	225	200	195			

(C. A. 1974)

[*Ans.* : 344, 332, 327, 320, 312, 300, 277, 259, 251, 247, 245, 238, 226, 218, 216, 215, 213]

3. The following series of observations is known to have a business cycle with a period of 4 years. Find the trend values by the moving average method selecting an appropriate period of the moving averages :

Year :	1951	1952	1953	1954	1955	1956	1957	1958	1959
Y :	506	620	1036	673	588	596	1116	738	663
Year :	1960	1961	1962	1963	1964	1955			
Y :	773	1189	818	745	845	1276			

(C. U., 1971)

[*Ans.* : 4-yearly m.a. : 719·0, 738·8, 758·8, 776·4, 793·9, 812·9, 831·6, 850·8, 871·0, 890·3, 910·1]

4. The following table shows the number of salesmen working for a certain concern :

Year :	1970	1971	1972	1973	1974
Number :	28	38	46	40	56

Use the method of least squares to fit a straight line and estimate the number of salesmen in 1975.

[*Ans.* : $y = 41{\cdot}6 + 5{\cdot}8t$; 59]

5. Fit a straight line trend equation by the method of least squares and estimate the value for 1969 :

Year :	1960	1961	1962	1963	1964	1965	1966	1967
Value :	380	400	650	720	690	600	870	930

(C. A. 1978)

[Ans. : $y = 655 + 35{\cdot}8t$; 1048·8]

6. Below are given the figures of production in thousand tons of a sugar factory :

Year :	1969	1970	1971	1972	1973	1974	1975
Production :	77	88	94	85	91	98	90

Fit a straight line by the method of least squares and show the trend values.

[Ans. : $y = 89 + 2t$; 83, 85, 87, 89, 91, 93, 95]

7. Fit a straight line trend by the method of least squares to the following data and obtain the trend value for the year 1972 :

Year :	1960	1961	1962	1963	1964	1965
Production (Lakh tons) :	3·6	3·8	4·4	4·7	5·6	7·3
Year :	1966	1967	1968	1969	1970	1971
Production (Lakh tons) :	7·1	7·6	7·7	9·0	9·0	10·1

[*Ans.* : $y = 6{\cdot}658 + {\cdot}599t$; 10·55]

8. Compute the trend values by the method of least squares from the data given below :

Year :	1962	1963	1964	1965	1966	1967	1968	1969
No. of sheeps (in Lakhs) :	56	55	51	47	42	38	35	32

[*Ans.* : $y = 44{\cdot}5 - 3{\cdot}71t$; origin : 1965·5]

9. Fit an equation of the type $y = a + bt + ct^2$ to the following data :

Year	1971	1972	1973	1974	1975	1976
Sales (in million tons)	100	105	115	100	112	118

[*Ans.* : $y = 107{\cdot}66 + 2{\cdot}74t + {\cdot}23t^2$; origin : middle of 1973-74]

10. Find the quarterly trend values from the following data by moving average method, using an appropriate period :

Quarterly Output (million tons)

Quarter/Year	1964	1965	1966
I	52	59	57
II	54	63	61
III	67	75	72
IV	55	65	60

(I. C. W. A., 71)

[*Ans.* : 4-quarter moving averages : 57·9, 59·9, 62·0, 64·2, 65·2, 64·8, 64·1, 63·1 (million tons)]

11. Find the trend values (mixed with cyclical movements, if any) from the following data by the method of moving averages :

Quarter/Year	1930	1931	1932	1933
I	29	40	47	45
II	37	42	51	49
III	43	55	63	60
IV	34	43	53	48

(I. C. W. A., 1968)

[*Ans.* : 4-quarterly moving averages : 37·1, 39·1, 41·2, 43·9, 45·9, 47·9, 50·0, 52·2, 53·2, 52·8, 52·1, 51·1]

12. Calculate the seasonal indices by the ratio to moving average method from the following data :

Year/Quarter	I	II	III	IV
1975	68	62	61	63
1976	65	58	66	61
1977	68	63	63	67

[*Ans.* : 105·30, 95·21, 100·97, 98·52]

13. Deseasonalise the following production data by the method of moving average :

Quarterly Output ('000 tons)

Quarter/Year	1970	1971	1972	1973
I	30	49	35	75
II	49	50	62	79
III	50	61	60	65
IV	35	20	25	70

(C. U., 1973)

[*Ans.* : 28·28, 38·70, 39·28, 57·74, 47·28, 39·70, 50·28, 42·74, 33·28, 51·70, 49·28, 47·74, 73·28, 68·70, 54·28, 92·74]

14. In a study of its sales, a motor company obtained the following least square equation :

$$y = 1600 + 200x$$

(origin at 1950, x units = 1 year, y = number of units sold annually)

The company has physical facilities to produce only 3,600 units a year and it believes that it is reasonable to assume that at least for the next decade the trend will continue as before :

(*a*) What is the average annual increase in the number of units sold ?

(*b*) Estimate the annual sales for 1965. How much in excess of the company's present physical capacity is this estimated value ?

(C. U., 1969)

[Ans. : (*a*) 200 units, (*b*) 4600 units, 1000 units]

15. The trend equation fitted to a series of sales data is given by

$$y = 1600 + 200x$$

(origin at 1950, x units = 1 year, y = number of units sold yearly)

The company has the production capacity of 3,600 units a year. Find by what year will the company's expected sales have equalled its present production capacity (assuming that at least for the next decade the trend will continue as before). (C. U., 1966)

[*Ans.* : year 1960]

16. Find the seasonal variations by the method of link relatives from the data given below :

Year/Quarter	I	II	III	IV
1975	65	58	56	61
1976	68	63	63	67
1977	70	59	56	52
1978	60	55	51	58

[*Ans.* : 109'3, 97'8, 93'9, 99'0]

17. Deseasonalise the following sales data and interpret them :

Quarter	Sales (Rs. '000)	Seasonal Indices
I	23'7	0'78
II	25'2	1'24
III	21'4	0'50
IV	65'4	1'48

(C. U., 1968)

[*Ans.* : 30'4, 20'3, 42'8, 44'2 ; Multiplicative model used]

18. The seasonal indices of the sale of ready-made garments of a particular type in a certain store are given below :

Quarter	*Seasonal Indices*
Jan.-March	98
April-June	89
July-Sept.	83
Oct.-Dec.	130

If the total sales in the 1st quarter of the year be worth Rs. 10,000, determine how much worth of garments of this type should be kept in store by the store to meet the demand in each of the remaining quarters.

[*Ans.* : Rs. 9081, Rs. 8469, Rs. 13265]

19. For the following data, obtain seasonal indices by any method that you consider suitable :

Passenger miles flown by domestic survices of U.K. Airlines (millions of passenger miles)

Year/Quarter	I	II	III	IV
1961	77	166	252	104
1962	99	191	287	123
1963	113	229	316	156
1964	152	255	357	180

(C. U., 1971)

[*Ans.* : Movirg average methcd, additive models −72'27, 23'84, 106'30, −57'87]

20. On the basis of quarterly sales (in Rs. Lakhs) of a certain commodity for the years 1961—'65, the following calculations were made :

Trend : $y = 25{\cdot}0 + 0{\cdot}6t$ with origin at 1st quarter of 1961, where t = time units (one quarter) and y = quarterly sales (Rs. Lakhs).

Seasonal Variations :

Quarter :	I	II	III	IV
Seasonal Indices :	90	95	110	105

Estimate the quarterly sales for the year 1962 (use multiplicative model).

(I. C. W. A., 1972)

[*Ans.* : 24'66, 26'60, 31'46, 30'66]

13

SET THEORY

Introduction

George Cantor (1845—1918), a German Mathematician, was the creator of set theory. On the basis of set theory he developed mathematical analysis. His work on the *Theory of sets* was accepted as fundamental contribution to mathematics.

Set

In our daily life we use phrases like a *bunch* of keys, a *set* of books, a tea *set*, a *pack* of cards, a *team* of players, a *class* of students, etc. Here the words bunch, set, pack, team, class—all indicate collections or aggregates. In mathematics also we are to deal with collections. Mathematicians use the word *set* for a well-defined collection of objects.

A **set** is a *well-defined* collection of distinct objects. Each object is said to be an *element* (or *member*) of the set.

We shall use capital letters A, B, C or X, Y, Z or P, Q, R to indicate sets and small letters a, b, c or x, y, z or p, q, r to denote elements of a set.

Symbol

The symbol $\in$ is used to denote 'is an element of' or 'is a member of' or 'belongs to'. Thus for $x \in$ A, read as x is an element of A or x belongs to A. Again for denoting 'not an element of' or 'does not belong to' we put a diagonal line through $\in$ thus $\notin$. So if y does not belong to A, we may write (using the above symbol), $y \notin$ A.

EXAMPLE :

If V is the set of all vowels, we can say $e \in$ V and $f \notin$ V.

Methods of Describing a Set.

There are two methods :

(1) ***Tabular Method*** (or Roster Method)

As mentioned before a set is denoted by capital letters, *i.e.*, A, B, X, Y, P, Q, etc. The general way of designating a set is writing all the elements (or members) within brackets ( ) or { } or [ ]. Thus set A may be written as A = {green, red, blue}.

The order of listing the elements is not important, so the same set A may be written again as A = {blue, green, red}. Further any element may be repeated any number of times without disturbing the set. The same set A can be taken as A = {blue, blue, green, red, red, red}.

For large number (finite) of elements we will use dots to represent the elements within the set. If A be a set of odd numbers upto 17, we may write (for convenient) A = {1, 3, 5, ···, 17}. Again, if A be a set of Prime Ministers, A = {Nehru, Sastri, Gandhi, Desai}.

(2) *Selector Method* (or Rule Method)

In this method, if all the elements of a set possess some common property, which distinguishes the same elements from other non-elements, then that property may be used to designate the set. For *example*, if x (an element of a set B) has the property having odd positive integer such that 3 is less than equal to x and x is less than equal to 17, then in short, we may write,

$$B = \{ x : x \text{ is an odd positive integer and } 3 < x < 17 \}$$

Similarly $C = \{ x : x$ is a day beginning with Monday $\}$.

Note : (1) ' : ' is used after x is to be read as 'such that'. In some cases '|' (a vertical line) is used, which is also to be read as 'such that'.

(2) If the elements do not possess the common property, then this method is not applicable.

Types of Sets

(1) *Finite Set :*

It is a set consisting of finite number of elements.

EXAMPLE :

$A = \{ 1, 2, 3, 4, 5 \}$

$B = \{ 2, 4, 6, \cdots\cdots, 50 \}$

$C = \{ x : x$ is number of students in a class $\}$.

(2) *Infinite Set :*

A set having an infinite number of elements.

EXAMPLE :

$A=\{1, 2, 3, \cdots\cdots\}$
$B=\{2, 4, 6, \cdots\cdots\}$
$C=\{x : x$ is number of stars in the sky $\}$.

(3) *Null or Empty or Void Set :*

It is a set having no element in it, and is usually denoted by ϕ (read as phi) or $\{\ \}$.

EXAMPLE :

The number of persons moving in air without any machine.
A set of positive numbers less than zero.
$A=\{x : x$ is a perfect square of an integer $5<x<8\}$.
$B=\{x : x$ is a negative integer whose square is $-1\}$.

(4) *Equal Set :*

Two sets A and B are said to be equal if all the elements of A belong to B and all the elements of B belong to A.

EXAMPLE :

$A=\{1, 2, 3, 4\}$, $B=\{3, 1, 2, 4\}$,
or $A=\{a, b, c,\}$, $B=\{a, a, a, c, c, b, b, b, b\}$.

Note : The order of writing the elements or repetition of elements does not change the nature of set.

Now, $A=B$ if and only if $\{x \in A \Leftrightarrow x \in B\}$

Let $A=\{x : x^2-7x+12=0\}$, $B=\{3, 4\}$, $C=\{3, 3, 4, 3, 4\}$

Then $A=B=C$, since elements which belong to any one set, also belong to the other two sets.

If $A=\{2, 3, 4\}$ $B=\{4, 2, 3\}$
$X=\{1, 3, 4\}$ $Y=\{2, 3, 5.\}$

Then $A=B$, and $X \neq Y$.

Again let $A=\{x : x$ is a letter in the word STRAND $\}$
$B=\{x : x$ is a letter in the word STANDARD $\}$
$C=\{x : x$ is a letter in the word STANDING $\}$

Here $A=B$, $B \neq C$, $A \neq C$.

(5) *Equivalent Set :*

If the total number of elements of one set is equal to the total number of elements of another set, then the two sets are said to be equivalent. It is not essential that the elements of the two sets should be same.

EXAMPLE :

$A=\{1, 2, 3, 4\}$ $B=\{b, a, l, l\}$

In A, there are 4 elements, 1, 2, 3, 4

In B, there are 4 elements, *b*, *a*, *l*, *l* (one to one correspondence)

Hence $A\equiv B$ (symbol $\equiv$ is used to denote equivalent set)

$A=\{3, 5, 8, 9\}$ $B=\{5, 5, 8, 9, 3, 8, 9\}$

$C=\{b, o, o, k\}$

Here $A=B$ and $A\equiv C$

(6) *Sub-set :*

If each element of the set A belongs to the set B, then A is said to be a sub-set of B. Symbolically, the relation is $A\subseteq B$ and read as A is a sub-set of B or A is contained in B or A is included in B.

It may be mentioned here that usually set A should be smaller than set B, may be equal also, but in no case A should be greater than B.

EXAMPLE :

If $B=\{1, 2, 3\}$, then the sub-sets of B are $\{1\}$, $\{2\}$, $\{3\}$, $\{1, 2\}$, $\{2, 3\}$, $\{1, 3\}$, $\{1, 2, 3\}$ and ϕ.

Note : (1) Every element of a set is an element of the same set, therefore every set is a sub-set of itself, *i.e.*, $A\subseteq A$.

(2) Null set contains no element, so all the elements of ϕ belong to every set, *i.e.*, $\phi\subseteq A$.

(3) It follows that every set has at least two sub-sets, *i.e.*, the null set and the set i tself.

(4) If $A\subseteq B$ and $B\subseteq C\Rightarrow A\subseteq C$.

(5) If $A\subseteq B$ and $B\subseteq A\Rightarrow A=B$.

(6) If $A\subseteq\phi$, than $A=\phi$.

(7) *Proper Sub-set :*

If each and every element of a set A are the elements of B and there exists atleast one element of B that does not belong to A, then the set A is said to be a *proper sub-set* of B (or B is called *super set* of A). Symbolically, we may write,

$A \subset B$ (read as A is proper sub-set of B).

And $B \subset A$ means A is a super set of B.

If $B = \{a, b, c\}$, then proper sub-sets are $\{a\}, \{b\}, \{c\}, \{a, b\}, \{b, c\}, \{a, c\}, \phi$.

(8) *Power Set :*

The family of all sub-sets of a given set A is known as *power set* and is denoted by P (A).

EXAMPLE :

(i) If $A = \{a\}$, then $P(A) = \{a\}, \phi$.

(ii) If $A = \{a, b\}$, then $P(A) = \{a\}, \{b\}, \{a, b\}, \phi$.

(iii) If $A = \{a, b, c\}$,

$P(A) = \{a\}, \{b\}, \{c\}, \{a, b\}, \{b, c\}, \{a, c\}, \{a, b, c\}, \phi$.

Thus when the number of elements of A is 1, then the number of sub-sets is 2, when the number of elements of A is 2, then the number of sub-sets is $4 = 2^2$ and when it is 3, the number of sub-sets is $8 = 2^3$. So, if A has n elements, P (A) will have 2^n sub-sets.

Universal Set

In mathematical discussion, generally we consider all the sets to be sub-sets of a fixed set, known as Universal set or Universe, denoted by U. An universal set may be finite or infinite.

EXAMPLE :

(i) A pack of cards may be taken as an universal set for a set of diamond or spade.

(ii) A set of integers is a Universal set for the set of even or odd numbers.

Venn diagram

John Venn, an English logician (1834—1923) invented this diagram to present pictorial representation. The diagrams display operations on sets. In a Venn diagram, we shall denote Universe U (or X) by a region enclosed within a rectangle and any sub-set of U will be shown by circle or closed curve.

Union of Sets

If A and B are two given sets then their union is the set of those elements that belong either to A or to B (or to both).

The union of A and B is denoted symbolically as A∪B (*read as* A union B or A cup B) In symbols,

$$A \cup B = \{ x : x \varepsilon A \text{ or } x \varepsilon B \}$$

EXAMPLE :

(i) Let $A = \{ 1, 2, 3, 4, 5 \}$, $B = \{ 2, 3, 5, 6, 7 \}$, $C = \{ 2, 4, 7, 8, 9 \}$

Then $A \cup B = \{ 1, 2, 3, 4, 5, 6, 7 \}$, and $B \cup A = \{ 1, 2, 3, 4, 5, 6, 7 \}$

$\therefore$ $A \cup B = B \cup A$ (commutative law)

Again $(A \cup B) \cup C = \{ 1, 2, 3, 4, 5, 6, 7, 8, 9 \}$

$(B \cup C) = \{ 2, 3, 4, 5, 6, 7, 8, 9 \}$

$A \cup (B \cup C) = \{ 1, 2, 3, 4, 5, 6, 7, 8, 9 \}$

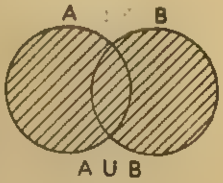
$\therefore$ $(A \cup B) \cup C = A \cup (B \cup C)$ (associative law)

(ii) If $A = \{ a, b, c, d \}$, $B = \{ 0 \}$, $C = \phi$

Then $A \cup B = \{ o, a, b, c, d \}$, $A \cup C = \{ a, b, c, d \} = A$, and $B \cup C = \{ 0 \}$

Union of sets may be illustrated more clearly by using *Venn Diagram* as follows :—

The shaded region indicates the union of A and B, *i.e.*, A∪B.,



Intersection of Sets

If A and B are two given sets, then their intersection is the set of those elements that belong to both A and B. and is denoted by A∩B (*read as* A intersection B or A cap B)

In symbols, $A \cap B = \{ x : x \varepsilon A \text{ and } x \varepsilon B \}$

EXAMPLE :

(i) For the same sets A, B; C given above in example (i),

$A \cap B = \{ 2, 3, 5 \}$, here the elements 2, 3, 5 belong both to A and B. And $B \cap A = \{ 2, 3, 5 \}$. $\therefore$ $A \cap B = B \cap A$ (commutative law)

$(A \cap B) \cap C = \{ 2 \}$

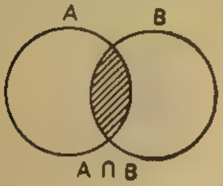
$(B \cap C) = \{ 2, 7 \}$, $A \cap (B \cap C) = \{ 2 \}$.

$\therefore$ $(A \cap B) \cap C = A \cap (B \cap C)$ (associative law)

(ii) For the sets A, B, C given in example (ii) above,

$A \cap B = \{ o \}$, $B \cap C = \phi$, $A \cap C = \phi$.

Intersection of two sets A and B is illustrated clearly by the *Venn Diagram* is as follows :—



The shaded portion represents the intersection of A and B, *i.e.*, $A \cap B$.

Disjoint Sets

Two sets A and B are said to be disjoint if their intersection is empty, *i.e.*, no element of A belongs to B.

EXAMPLE :

$A = \{ 1, 3, 5 \}$, $B = \{ 2, 4 \}$,

$A \cap B = \phi$. $\therefore$ A and B are disjoint sets.

Difference of Two Sets

If A and B are two sets, then the set containing all those elements of A which do not belong to B, is known as difference of two sets, and is denoted by the symbol $A \sim B$ or $A - B$ (*read* A difference B). Now, $A \sim B$ is said to be obtained by substracting B from A.

In symbols, $A \sim B = \{ x : x \in A \text{ and } x \notin B \}$.

EXAMPLE :

(i) If $A = \{ 1, 2, 3, 4, 5 \}$, $B = \{ 3, 5, 6, 7 \}$, then $A \sim B = \{ 1, 2, 4 \}$

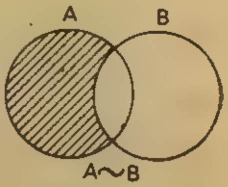
(ii) If $A = \{ x : x \text{ is an integer and } 1 < x < 12 \}$,

$B = \{ x : x \text{ is an integer and } 7 < x < 14 \}$,

then $A \sim B = \{ x : x \text{ is an integer and } 1 < x < 6 \}$

$A \sim B$ is represented by a *Venn Diagram* as follows :

The shaded portion represents $A \sim B$.



Complement of a Set

Let U be the universal set and A be its sub-set. Then the complement set of A in relation to U is that set whose elements belong to U and not to A. This is denoted by A' $(= U \sim A)$ or A^c or $\bar{A}$.

In symbols, $A' = \{x : x \in U \text{ and } x \notin A\}$. We may also write $A' = \{x : x \notin A\}$

Remarks :

1. The union of any set A and its complement A′ is the universal set, *i.e.*, $A \cup A' = U$.

2. The intersection of any set and its complement A′ is the null set, *i.e.*, $A \cap A' = \phi$.

EXAMPLE :

$$U = \{1, 2, 3, \cdots\cdots, 10\},\ A = \{2, 4, 7\}$$
$$A'(= U \sim A) = \{1, 3, 5, 6, 8, 9, 10\}$$

Now $A \cup A' = \{1, 2, 3, \cdots\cdots 10\} = U$, $A \cup A' = \phi$

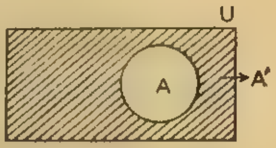
Again $(A')' = \{2, 4, 7\} = A$ (*i.e.*, complement of the complement of A is equal to A itself).

$U' = \phi$ (*i.e.*, complement of a universal set is empty).

Again the complement of an empty set is a universal set, *i.e.*, $\phi' = U$.

It $A \subset B$ then $B' \subset A'$ for sets A and B.

Complement of A is represented by the shaded region.



Symmetric Difference

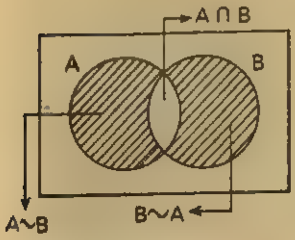
For the two sets A and B, the symmetric difference is $(A \sim B) \cup (B \sim A)$ and is denoted by $A \triangle B$ (*read* as A symmetric difference B).

EXAMPLE :

Let $A = \{1, 2, 3, 4, 8\}$, $B = \{2, 4, 6, 7\}$

Now, $A \sim B = \{1, 3, 8\}$, $B \sim A = \{6, 7\}$

$\therefore\ A \triangle B = \{1, 3, 8\} \cup \{6, 7\} = \{1, 3, 6, 7, 8\}$



By *Venn Diagram* :

$A \triangle B$ is represented by shaded region.

It is clear that $A \triangle B$ denotes the set of all those elements that belong to A and B except those which do not belong to A and B both, *i.e.*, it is the set of elements which belongs to A or B but not to both.

Difference between ϕ, $\{0\}$ and $\{\phi\}$.

ϕ is a null set

$\{0\}$ is a singleton whose only element is zero.

$\{\phi\}$ is also a singleton whose only element is a null set.

Worked out Examples

1. Rewrite the following examples using set notations :

(i) First ten even natural numbers.

(ii) Set of days of a week.

(iii) Set of months in a year which have 30 days.

(iv) The numbers 3, 6, 9, 12, 15

(v) The letters *m, a, t, h, e, m, a, t, i, c, s*

SOLUTION :

(i) $A = \{ 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 \}$ (Tabular method)
$= \{ x : x$ is an even integer and $2 \leqslant x \leqslant 20 \}$ (Selector method)

(ii) $A = \{$ Sunday, Monday, ... Saturday $\}$ (Tabular)
$= \{ x : x$ is a day in a week $\}$ (Selector)

(iii) $A = \{$ April, June, September, November $\}$ (Tabular)
$= \{ x : x$ is a month of 30 days $\}$ (Selector)

(iv) $A = \{ x : x$ is a positive number multiple of 3 and $3 \leqslant x \leqslant 15 \}$

(v) $A = \{ x : x$ is a letter in the word mathematics $\}$

2. Represent the following sets in selector method :

(i) all numbers less than 15

(ii) all even numbers

(iii) all real numbers in *closed* interval $\{ 1, 11 \}$

(iv) all real numbers in *open* interval $\{ -2, 3 \}$

SOLUTION :

Taking A be the set of numbers in every case :

(i) $\{ x : x \varepsilon A$ and $x < 15 \}$

(ii) $\{ x : x \varepsilon A$ and x is a multiple of $2 \}$

(iii) $\{ x : x \varepsilon A$ and $1 \leqslant x \leqslant 11 \}$

(iv) $\{ x : x \varepsilon A$ and $-2 < x < 3 \}$

3. Given $A = \{$the odd numbers between 2 and 10$\}$, $B = \{3, 4, 5\}$, $C = \{$all integers less than 26 which are perfect square$\}$, $D = \{1, x, x^2\}$, $E = \phi$, $F = \{1\}$, $G = \{$all the digits in the product of 5 and 47$\}$, $H = \{$all numbers between 90 and 110 which are multiple of 7$\}$, $I = \{1, 2, 3, 4, 5, 6\}$, $K = \{17\}$.

For each of the following sets, write down the set given above which is equal to :

(i) $\{1, 4, 9, 16, 25\}$ (ii) $\{x : x \varepsilon I$ and $2x - 7 = 0\}$

(iii) $\{$all prime factors of 30$\}$ (iv) $\{x : x \varepsilon N$ and $1 \leqslant x < 7\}$

(v) $\{3, 5, 7, 9\}$ (vi) $\{x^2, 1, x\}$

(vii) {all prime numbers greater than 13 and less than 19}

(viii) {all numbers by adding a pair of the numbers from 1, 2, 3}

(ix) $\{x : x \varepsilon \text{N and } x = 1\}$ (x) $\{91, 98, 105\}$

SOLUTION :

(i) C, (ii) E, (iii) G {prime factors of 30 are 2, 3, 5, again $5 \times 47 = 235$, digits are 2, 3, 5}, (iv) I, (v) A, (vi) D, (vii) K, (viii) B {adding each pair, we find 3, 4, 5}, (ix) F, (x) H.

4. (i) Is the set $A = \{x : x < x\}$ a null ?

(ii) Is the set $B = \{x : x + 4 = 4\}$ a null ?

(iii) Is the set $C = \{x : x$ is a positive number less than zero$\}$ a null ?

SOLUTION :

(i) Null, as there exists no number less than itself.

(ii) Not null, the set has an element zero.

(iii) Null, as there exists no positive number less than zero.

5. State with reasons whether each of the following statement is true or false :

(i) $\{1\} \varepsilon \{1, 2, 3\}$ (ii) $1 \varepsilon \{1, 2, 3\}$

(iii) $\{1\} \subset \{1, 2, 3\}$ (iv) $1 \subset \{1, 2, 3\}$

(v) $\{1, 2\} \varepsilon \{1, 2, 4\}$ (vi) $\{1, 2\} \subset \{1, 2, 3\}$

(vii) $\{1, 2, 3\} = \{2, 3, 4\}$ (viii) $\{1, 2, 3\} = \{3, 2, 3, 1, 2, 1\}$

(ix) $\{1, 2, 3\} \varepsilon \{1, 2, 3\}$ (x) $\{1, 2, 3\} \subset \{3, 1, 2\}$

(xi) $\phi \varepsilon \{1, 2, 3\}$ (xii) $4 \notin \{1, 2, 3\}$

SOLUTION :

(i) False, $\{1\}$ is a singleton and not an element of $\{1, 2, 3\}$

(ii) True, since 1 is an element and belongs of $\{1, 2, 3\}$

(iii) True, $\{1\}$ is a proper sub-set of $\{1, 2, 3\}$

(iv) False, an element can not be a sub-set of a set.

(v) $\{1, 2\}$ is not an element but a sub-set of $\{1, 2, 3\}$, so it is false.

(vi) True, $\{1, 2\}$ is a proper sub-set of $\{1, 2, 3\}$

(vii) False, as $1 \notin \{2, 3, 4\}$ and $4 \notin \{1, 2, 3\}$

(viii) True, since both sets contain same element.

(ix) False, a set does not belong to the same set.

(x) False, as $\{1, 2, 3\}$ is not a proper sub-set of $\{3, 1, 2\}$

(xi) False, null set is not an element of $\{1, 2, 3\}$

(xii) True.

Few Properties

On Union of Sets :

Some properties for sets A, B and C are—

(1) $A \subseteq (A \cup B)$ and $B \subseteq (A \cup B)$

(2) $A \cup A = A$

(3) $A \cup \phi = A$ (identity property)

(4) $A \cup B = \phi \Rightarrow A = \phi, B = \phi$

(5) $A \cup B = B \cup A$ (commutative property)

(6) $(A \cup B) \cup C = A \cup (B \cup C)$ (associative property)

PROOF OF COMMUTATIVE LAW : $A \cup B = B \cup A$

We are to show (i) $A \cup B \subseteq B \cup A$

(ii) $B \cup A \subseteq A \cup B$

(i) Let any element $x \varepsilon A \cup B$ then

$$x \varepsilon A \cup B \Rightarrow x \varepsilon A \text{ or } x \varepsilon B$$
$$\Rightarrow x \varepsilon B \text{ or } x \varepsilon A$$

$\Rightarrow x \varepsilon (B \cup A)$ which shows that any element of $A \cup B$ is also an element of $B \cup A$.

$$\therefore \quad A \cup B \subseteq B \cup A \quad \cdots \quad \cdots \quad (1)$$

(ii) Let any element y belong to $B \cup A$, then

$$y \varepsilon B \cup A \Rightarrow y \varepsilon B \text{ or } y \varepsilon A$$
$$\Rightarrow y \varepsilon A \text{ or } y \varepsilon B$$

$\Rightarrow y \varepsilon (A \cup B)$ which shows that any element of $B \cup A$ is also an element of $A \cup B$.

$$\therefore \quad B \cup A \subseteq A \cup B \quad \cdots \quad \cdots \quad (2)$$

From (1) and (2) we find $A \cup B = B \cup A$.

PROOF OF ASSOCIATIVE PROPERTY : $(A \cup B) \cup C = A \cup (B \cup C)$

We are to show (i) $(A \cup B) \cup C \subseteq A \cup (B \cup C)$

(ii) $A \cup (B \cup C) \subseteq (A \cup B) \cup C$

(i) Let x be any element of $(A \cup B) \cup C$. Then

$$x \varepsilon (A \cup B) \cup C \Rightarrow x \varepsilon (A \cup B) \text{ or } x \varepsilon C$$
$$\Rightarrow (x \varepsilon A \text{ or } x \varepsilon B) \text{ or } x \varepsilon C$$
$$\Rightarrow x \varepsilon A \text{ or } (x \varepsilon B \text{ or } x \varepsilon C)$$
$$\Rightarrow x \varepsilon A \text{ or } x\, (B \cup C)$$
$$\Rightarrow x \varepsilon A \cup (B \cup C)$$

which shows that every element of $(A\cup B)\cup C$ is also an element of $A\cup(B\cup C)$

$$\therefore\ (A\cup B)\cup C\subseteq A\cup(B\cup C)\ \cdots\ \cdots\ (1)$$

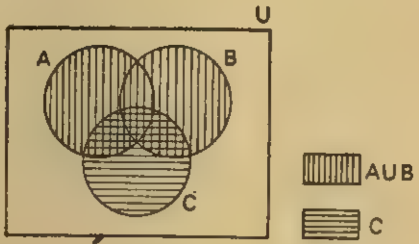
(ii) Let y be any element of $A\cup(B\cup C)$. Then

$$\begin{aligned} y\in A\cup(B\cup C) &\Rightarrow y\in A \text{ or } y\in(B\cup C)\\ &\Rightarrow y\in A \text{ or } (y\in B \text{ or } y\in C)\\ &\Rightarrow (y\in A \text{ or } y\in B) \text{ or } y\in C\\ &\Rightarrow y\in(A\cup B) \text{ or } y\in C\\ &\Rightarrow y\in(A\cup B)\cup C \end{aligned}$$

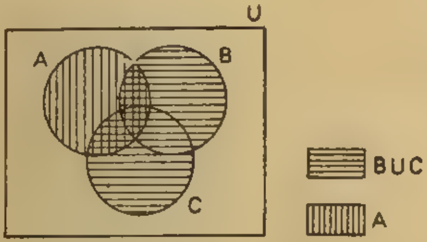
So we have, $A\cup(B\cup C)\subseteq(A\cup B)\cup C\ \cdots\ \cdots\ (2)$

From (1), (2) we can say, by the definition of equality of sets, $(A\cup B)\cup C=A\cup(B\cup C)$.

By *Venn Diagram* :



TOTAL SHADED REGION
(A∪B) ∪C



TOTAL SHADED REGION
A∪(B∪C)

On Intersection of Sets :

Some properties for sets A, B and C are—

(1) $A\cap B\subseteq A$ and $A\cap B\subseteq B$

(2) $A\cap\phi=\phi$

(3) $A \cap A = A$

(4) $A \cap B = B \cap A$ (commutative property)

(5) $(A \cap B) \cap C = A \cap (B \cap C)$ (associative property)

(6) $A \subseteq B$ then $A \cap B = A$ and if $B \subseteq A$ then $A \cap B = B$.

(7) If $A \subseteq B$ and $B \subseteq C$, then $A \subseteq (B \cap C)$.

Proof of commutative property, *i.e.*, $A \cap B = B \cap A$ is similar to $A \cup B = B \cup A$ which is shown above and is left to the students as an exercise.

Proof of Associative Property : $(A \cap B) \cap C = A \cap (B \cap C)$

[C.A. Entr., May 76]

We are to show (i) $(A \cap B) \cap C \subseteq A \cap (B \cap C)$

(ii) $A \cap (B \cap C) \subseteq (A \cap B) \cap C$

(i) Let x be any element of $(A \cap B) \cap C$. Then

$$\begin{aligned} x \,\varepsilon\, (A \cup B) \cap C &\Rightarrow x \,\varepsilon\, (A \cap B) \text{ and } x \,\varepsilon\, C \\ &\Rightarrow (x \,\varepsilon\, A \text{ and } x \,\varepsilon\, B) \text{ and } x \,\varepsilon\, C \\ &\Rightarrow x \,\varepsilon\, A \text{ and } (x \,\varepsilon\, B \text{ and } x \,\varepsilon\, C) \\ &\Rightarrow x \,\varepsilon\, A \text{ and } x \,\varepsilon\, (B \cap C) \\ &\Rightarrow x \,\varepsilon\, A \cap (B \cap C) \end{aligned}$$

Thus every element x of $(A \cap B) \cap C$ is also an element of $A \cap (B \cap C)$

$$\therefore \quad (A \cap B) \cap C \subseteq A \cap (B \cap C) \quad \cdots \quad \cdots \quad (1)$$

(ii) Let y be any element of $A \cap (B \cap C)$. Then

$$\begin{aligned} y \,\varepsilon\, A \cap (B \cap C) &\Rightarrow y \,\varepsilon\, A \text{ and } y \,\varepsilon\, (B \cap C) \\ &\Rightarrow y \,\varepsilon\, A \text{ and } (y \,\varepsilon\, B \text{ and } y \,\varepsilon\, C) \\ &\Rightarrow (y \,\varepsilon\, A \text{ and } y \,\varepsilon\, B) \text{ and } y \,\varepsilon\, C \\ &\Rightarrow y \,\varepsilon\, (A \cap B) \text{ and } y \,\varepsilon\, C \\ &\Rightarrow y \,\varepsilon\, (A \cap B) \cap C \end{aligned}$$

The every element y of $A \cap (B \cap C)$ is also an element of $(A \cap B) \cap C$

$$\therefore \quad A \cap (B \cap C) \subseteq (A \cap B) \cap C \quad \cdots \quad \cdots \quad (2)$$

From (1) and (2), we have $A \cap (B \cap C) = (A \cap B) \cap C$

Note : Property (6) and (7) may be illustrated by Venn Diagram which is left to students as an exercise.

On Union and Intersection of Sets (Union distributes over Intersection):

DISTRIBUTIVE LAW :

1. Let A, B and C are any three sets, prove that $A\cup(B\cap C)=(A\cup B)\cap(A\cup C)$. Also verify the result by Venn Diagram. [C.A. Inter, May '75, Nov. '76]

Here we are to show (i) $A\cup(B\cap C)\subseteq(A\cup B)\cap(A\cup C)$

(ii) $(A\cup B)\cap(A\cup C)\subseteq A\cup(B\cap C)$

(i) Let x be any arbitrary element of $A\cup(B\cap C)$. Then

$$\begin{aligned}x\;\varepsilon\;A\cup(B\cap C) &\Rightarrow x\;\varepsilon\;A \text{ or } x\;\varepsilon\;(B\cap C)\\ &\Rightarrow x\;\varepsilon\;A \text{ or } (x\;\varepsilon\;B \text{ and } x\;\varepsilon\;C)\\ &\Rightarrow (x\;\varepsilon\;A \text{ or } x\;\varepsilon\;B) \text{ and } (x\;\varepsilon\;A \text{ or } x\;\varepsilon\;C)\\ &\Rightarrow x\;\varepsilon\;(A\cup B) \text{ and } x\;\varepsilon\;(A\cup C)\\ &\Rightarrow x\;\varepsilon\;(A\cup B)\cap(A\cup C)\end{aligned}$$

Thus every element x of $A\cup(B\cap C)$ also belongs to

$(A\cup B)\cap(A\cup C)$

$\therefore\quad A\cup(B\cap C)\subseteq(A\cup B)\cap(A\cup C)\quad\cdots\quad\cdots\quad(1)$

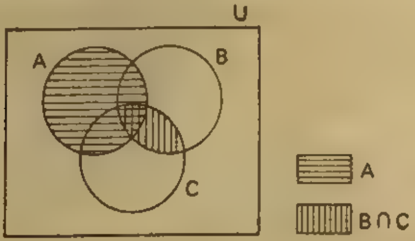
(ii) Let y be any element of $(A\cup B)\cap(A\cup C)$. Then

$$\begin{aligned}y\;\varepsilon\;(A\cup B)\cap(A\cup C) &\Rightarrow y\;\varepsilon\;(A\cup B) \text{ and } y\;\varepsilon\;(A\cup C)\\ &\Rightarrow (y\;\varepsilon\;A \text{ or } y\;\varepsilon\;B) \text{ and } (y\;\varepsilon\;A \text{ or } y\;\varepsilon\;C)\\ &\Rightarrow y\;\varepsilon\;A \text{ or } (y\;\varepsilon\;B \text{ and } y\;\varepsilon\;C)\\ &\Rightarrow y\;\varepsilon\;A \text{ or } y\;\varepsilon\;(B\cap C)\\ &\Rightarrow y\;\varepsilon\;A\cup(B\cap C)\end{aligned}$$

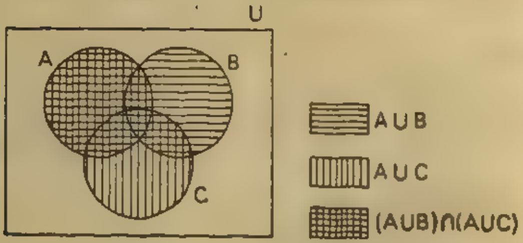
In the same way, $(A\cup B)\cap(A\cup C)\subseteq A\cup(B\cap C)\quad\cdots\quad\cdots\quad(2)$

From (1) and (2), we can say, $A\cup(B\cap C)=(A\cup B)\cap(A\cup C)$

By *Venn Diagram* :



TOTAL SHADED REGION
AU(B∩C)



(Intersection distributes over union)

2. Let A, B and C are any three sets, prove that $A\cap(B\cup C)=(A\cap B)\cup(A\cap C)$

As before, we are to show (i) $A\cap(B\cup C)\subseteq(A\cap B)\cup(A\cap C)$

(ii) $(A\cap B)\cup(A\cap C)\subseteq A\cap(B\cup C)$

(i) Let $x\in A\cap(B\cup C)\Rightarrow x\in A$ and $x\in(B\cup C)$

$\Rightarrow x\in A$ and ($x\in B$ or $x\in C$)

$\Rightarrow$ ($x\in A$ and $x\in B$) or ($x\in A$ and $x\in C$)

$\Rightarrow x\in(A\cap B)\cup(A\cap C)$

Thus $A\cap(B\cup C)\subseteq(A\cap B)\cup(A\cap C)$ ... ... (1)

(ii) Let $y(A\cap B)\cup(A\cap C)\Rightarrow y\in(A\cap B)$ or $y\in(A\cap C)$

$\Rightarrow$ ($y\in A$ and $y\in B$) or ($y\in A$ and $y\in C$)

$\Rightarrow y\in A$ and ($y\in B$ or $y\in C$)

$\Rightarrow y\in A\cap(B\cup C)$

Thus $(A\cap B)\cup(A\cap C)\subseteq A\cap(B\cup C)$ ... ... (2)

From (1) and (2), we find, $A\cap(B\cup C)=(A\cap B)\cup(A\cap C)$

Note : The student may verify the result by Venn Diagram.

Duality

Union and intersection are dual operations to each other. If we can establish the validity of one law having $\cup$ and $\cap$ as operations, then its dual will be also true, replacing $\cup$ by $\cap$ and $\cap$ by $\cup$.

De Morgan's Laws :

1. **Complement of a union is the intersection of complements.**

For any two sets A and B. Prove that $(A\cup B)'=A'\cap B'$

[C.A. Entrance, May '76]

We are to show (i) $(A\cup B)'\subseteq A'\cap B'$

(ii) $A'\cap B'\subseteq(A\cup B)'$

(i) Let $x \varepsilon (A\cup B)' \Rightarrow x \notin (A\cup B)$

$\Rightarrow x \notin A$ and $x \notin B$

$\Rightarrow x \in A'$ and $x \in B'$

$\Rightarrow x \in A'\cap B'$

Thus every element of $(A\cup B)'$ is also a member of $A'\cap B'$

$\therefore\ (A\cup B)'\subseteq A'\cap B'$ ... ... (1)

(ii) Let $y \varepsilon A'\cap B' \Rightarrow y \varepsilon A'$ and $y \varepsilon B'$

$\Rightarrow y \notin A$ and $y \notin B$

$\Rightarrow y \notin (A\cup B)$

$\Rightarrow y \varepsilon (A\cup B)'$

Thus every element belonging to $A'\cap B'$ also belongs to $(A\cup B)'$

$\therefore\ A'\cap B'\subseteq(A\cup B)'$ ... ... (2)

From (1) and (2), we find $(A\cup B)'=A'\cap B'$

Note : It may be noted that when union is followed by $\notin$ it changes to intersection and vice versa.

2. **Complement of an intersection is the union of the complements, *i.e.*, $(A\cap B)'=A'\cup B'$.**

(The proof is left to the student. The student may follow the above method.)

Some Important Results :

1. $(A\cup B)\cap(A\cup B')=A.$

For, $(A\cup B)\cap(A\cup B')=A\cup(B\cap B')=A\cup\phi=A.$

2. $(A\cap B)\cup(A\cap B')=A.$

For, $(A\cap B)\cup(A\cap B')=A\cap(B\cup B')=A\cap U=A.$

De Morgon's Laws on Difference of Set :

1. If A, B and C are three sets, then show that

$$A\sim(B\cup C)=(A\sim B)\cap(A\sim C)$$

[C.A. Entrance, May '75, Inter, May '76]

Let $x \in A\sim(B\cup C) \Rightarrow x \in A$ and $x\notin (B\cup C)$

$\Rightarrow x \in A$ and ($x\notin B$ and $x \notin C$)

$\Rightarrow$ ($x \in A$ and $x \notin B$) and ($x \in A$ and $x \notin C$)

$\Rightarrow x \varepsilon (A\sim B)$ and $x \varepsilon (A\sim C)$

$\Rightarrow x \varepsilon (A\sim B)\cap(A\sim C)$.

Thus, $A \sim (B \cup C) \subseteq (A \sim B) \cap (A \sim C)$

Again, let $y \varepsilon (A \sim B) \cap (A \sim C) \Rightarrow y \varepsilon (A \sim B)$ and $y \varepsilon (A \sim C)$

$\Rightarrow (y \in A$ and $y \notin B)$ and $(y \in A$ and $y \notin C)$

$\Rightarrow y \in A$ and $(y \notin B$ and $y \notin C)$

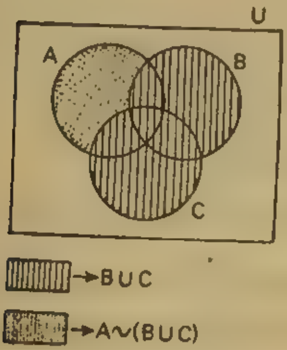
$\Rightarrow y \in A$ and $y \notin (B \cup C)$

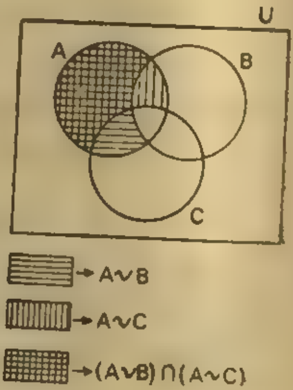
$\Rightarrow y \varepsilon A \sim (B \cup C)$

Thus, $(A \sim B) \cap (A \sim C) \subseteq A \sim (B \cup C)$

$\therefore \quad A \sim (B \cup C) = (A \sim B) \cap (A \sim C)$

By *Venn Diagram* :





2. If A, B and C are three sets, then show that

$$A \sim (B \cap C) = (A \sim B) \cup (A \sim C)$$

Let $x \in A \sim (B \cap C) \Rightarrow x \in A$ and $x \notin (B \cap C)$

$\Rightarrow x \in A$ and $(x \notin B$ or $x \notin C)$

$\Rightarrow (x \in A$ and $x \notin B)$ or $(x \in A$ and $x \notin C)$

$\Rightarrow x \varepsilon (A \sim B)$ or $x \varepsilon (A \sim C)$

$\Rightarrow x \varepsilon (A \sim B) \cup (A \sim C)$

i.e., $A \sim (B \cap C) \subseteq (A \sim B) \cup (A \sim C)$

Again, let $y \varepsilon (A \sim B) \cup (A \sim C) \Rightarrow y \varepsilon (A \sim B)$ or $y \varepsilon (A \sim C)$

$\Rightarrow (y \in A$ and $y \notin B)$ or $(y \in A$ and $y \notin C)$

$\Rightarrow y \in A$ and $(y \notin B$ or $y \notin C)$

$\Rightarrow y \in A$ and $y \notin (B \cap C)$

$\Rightarrow y \varepsilon A \sim (B \cap C)$

i.e., $(A \sim B) \cup (A \sim C) \subseteq A \sim (B \cap C)$

$\therefore \quad A \sim (B \cap C) = (A \sim B) \cup (A \sim C)$

This result may also be verified by *Venn Diagram* as before. The student may do it now.

Miscellaneous Results (on Union, Intersection and Difference).

1. Prove that $A \sim B = A \cap B'$

Let $x \in A \sim B \Rightarrow x \in A$ and $x \notin B$
$\Rightarrow x \varepsilon A$ and $x \varepsilon B'$
$\Rightarrow x \varepsilon A \cap B'$

Thus, $A \sim B \subsetneq A \cap B'$

Again, let $y \varepsilon A \cap B' \Rightarrow y \varepsilon A$ and $y \varepsilon B'$
$\Rightarrow y \in A$ and $y \notin B$
$\Rightarrow y \varepsilon A \sim B$

Thus, $A \cap B' \subseteq A \sim B$

$\therefore \quad A \sim B = A \cap B'$

Note : The student may now try to show $B \sim A = B \cap A'$.

2. Prove that $(A \sim B) \cap B = \phi$.

Let x be at least one element belongs to $(A \sim B) \cap B$
then $x \varepsilon (A \sim B) \cap B \Rightarrow x \varepsilon (A \sim B)$ and $x \varepsilon B$
$\Rightarrow (x \in A$ and $x \notin B)$ and $x \in B$
$\Rightarrow x \in A$ and $x \notin B$ and $x \in B$, which is absurd,
since $x \in B$ and $x \notin B$ cannot hold simultaneously.

$\therefore \quad (A \sim B) \cap B = \phi$.

3. Prove that $A \cap (B \sim C) = (A \cap B) \sim C$

We are to show $A \cap (B \sim C) \subsetneq (A \cap B) \sim C$ and
$(A \cap B) \sim C \subsetneq A \cap (B \sim C)$

For the first one, let $x \varepsilon A \cap (B \sim C)$
$\Rightarrow x \varepsilon A$ and $x \varepsilon (B \sim C)$
$\Rightarrow x \in A$ and $(x \in B$ and $x \notin C)$
$\Rightarrow (x \in A$ and $x \in B)$ and $x \notin C$
$\Rightarrow x \in A \cap B$ and $x \notin C$
$\Rightarrow x \varepsilon (A \cap B) \sim C$

Thus, $A \cap (B \sim C) \subsetneq (A \cap B) \sim C$

For the second part, let $y \epsilon (A \cap B) \sim C$

$\Rightarrow y \in A \cap B$ and $y \notin C$

$\Rightarrow (y \in A$ and $y \in B)$ and $y \notin C$

$\Rightarrow y \in A$ and $(y \in B$ and $y \notin C)$

$\Rightarrow y \epsilon A$ and $y \epsilon (B \sim C)$

$\Rightarrow y \epsilon A \cap (B \sim C)$

Thus, $(A \cap B) \sim C \subseteq A \cap (B \sim C)$

$\therefore \quad A \cap (B \sim C) = (A \cap B) \sim C$

4. Prove that $A \cap (B \sim C) = (A \cap B) \sim (A \cap C)$.

Let, $x \epsilon A \cap (B \sim C)$

$\Rightarrow x \epsilon A$ and $x \epsilon (B \sim C)$

$\Rightarrow x \in A$ and $(x \in B$ and $x \notin C)$

$\Rightarrow (x \in A$ and $x \in B)$ and $x \notin C$

$\Rightarrow (x \epsilon A$ and $x \epsilon B)$ and $x \epsilon C'$

$\Rightarrow x \epsilon (A \cap B) \cap C'$

$\Rightarrow x \epsilon \phi \cup \{(A \cap B) \cap C'\}$

$\Rightarrow x \epsilon (\phi \cap B) \cup \{(A \cap B) \cap C'\}$

$\Rightarrow x \epsilon \{(A \cap A') \cap B\} \cup \{(A \cap B) \cap C'\}$

$\Rightarrow x \epsilon \{(A \cap B) \cap A'\} \cup \{(A \cap B) \cap C'\}$

$\Rightarrow x \epsilon (A \cap B) \cap (A' \cup C')$

$\Rightarrow x \epsilon (A \cap B) \cap (A \cap C)'$

$\Rightarrow x \epsilon (A \cap B) \sim (A \cap C)$

i.e., $A \cap (B \sim C) \subseteq (A \cap B) \sim (A \cap C)$.

Again let, $y \epsilon (A \cap B) \sim (A \cap C)$

$\Rightarrow y \in (A \cap B)$ and $y \notin (A \cap C)$

$\Rightarrow y \epsilon (A \cap B)$ and $y \epsilon (A \cap C)'$

$\Rightarrow y \epsilon (A \cap B)$ and $y \epsilon (A' \cup C')$

$\Rightarrow y \epsilon \{(A \cap B) \cap A'\} \cup \{(A \cap B) \cap C'\}$

$\Rightarrow y \epsilon \{(A \cap A') \cap B\} \cup \{(A \cup B) \cap C'\}$

$\Rightarrow y \epsilon (\phi \cap B) \cup \{(A \cap B) \cap C'\}$

$\Rightarrow y \epsilon \phi \cup \{(A \cap B) \cap C'\}$

$\Rightarrow y \epsilon (A \cap B) \cap C'$

$\Rightarrow y \epsilon A \cap (B \cap C')$

$\Rightarrow y \epsilon A \cap (B \sim C)$

i.e., $(A \cap B) \sim (A \cap C) \subseteq A \cap (B \sim C)$.

Thus, $A \cap (B \sim C) = (A \cap B) \sim (A \cap C)$.

5. Prove that $(A \sim B) \cup (B \sim A) = (A \cup B) \sim (A \cap B)$

Let $x \varepsilon (A \sim B) \cup (B \sim A) \Rightarrow x \varepsilon (A \sim B)$ or $x \varepsilon (B \sim A)$

$\Rightarrow (x \in A$ and $x \notin B)$ or $(x \in B$ and $x \notin A)$

$\Rightarrow x \in A$ or $x \in B$, but $\notin$ both A and B

$\Rightarrow x \in A \cup B$, but $\notin A \cap B$

$\Rightarrow x \varepsilon (A \cup B) \sim (A \cap B)$

Thus $(A \sim B) \cup (B \sim A) \subseteq (A \cup B) \sim (A \cap B)$

Similarly, $(A \cup B) \sim (A \cap B) \subseteq (A \sim B) \cup (B \sim A)$

$\therefore \quad (A \sim B) \cup (B \sim A) = (A \cup B) \sim (A \cap B)$

Note : $A \triangle B$ (A Symmetric difference B) $= (A \sim B) \cup (B \sim A)$.

6. For any two sets A, B, show that

$$A \cup B = (A \sim B) \cup B$$

Let $x \varepsilon A \cup B \Rightarrow x \varepsilon A$ or $x \varepsilon B$

$\Rightarrow x \varepsilon B$ or $x \varepsilon A$

$\Rightarrow (x \in B$ or $x \in A)$ and $(x \in B$ or $x \notin B)$

(step may be noted)

$\Rightarrow x \in B$ or $(x \in A$ and $x \notin B)$

$\Rightarrow x \varepsilon B$ or $x \varepsilon (A \sim B)$

$\Rightarrow x \varepsilon (A \sim B)$ or $x \varepsilon B$

$\Rightarrow x \varepsilon (A \sim B) \cup B$

i.e., $A \cup B \subseteq (A \sim B) \cup B$

Again, let $y \varepsilon (A \sim B) \cup B \Rightarrow y \varepsilon (A \sim B)$ or $y \varepsilon B$

$\Rightarrow y \varepsilon B$ or $y \varepsilon (A \sim B)$

$\Rightarrow y \in B$ or $(y \in A$ and $y \notin B)$

$\Rightarrow (y \varepsilon B$ or $y \varepsilon A)$ and $(y \varepsilon B$ or $y \varepsilon B)$

$\Rightarrow y \varepsilon B$ or $y \varepsilon A$

$\Rightarrow y \varepsilon A$ or $y \varepsilon B$

$\Rightarrow y \varepsilon (A \cup B)$

i.e., $(A \sim B) \cup B \subseteq A \cup B$

$\therefore \quad A \cup B = (A \sim B) \cup B$

7. Prove that $A \subset B$ if and only if $A \cup B = B$

Let us suppose $A \subset B$, we will show $A \cup B = B$

$$
\begin{aligned}
\text{Let } x \in A\cup B &\Rightarrow x \in A \text{ or } x \in B\\
&\Rightarrow x \in B \text{ or } x \in B \text{ (as } A\subset B)\\
&\Rightarrow x \in B\\
\text{i.e., } & A\cup B \subseteq B\\
\text{Again, } y \in B &\Rightarrow y \in A\cup B\\
\text{i.e., } & B\subsetneq A\cup B\\
\therefore\ & A\cup B = B.
\end{aligned}
$$

Algebra of Sets

The algebra of sets deals with certain fundamental laws or properties, governing operations on sets. These are like fundamental laws of addition and multiplication in ordinary algebra of numbers. Only in certain stages there are differences, which are discussed below :

(i) ***Commutative law*** **:**

For real numbers, addition and multiplication are commutative,

i.e., $a+b=b+a$ and $a\times b=b\times a$.

Union and Intersection of sets are also commutative,

i.e., $A\cup B=B\cup A$ and $A\cap B=B\cap A$.

(ii) ***Associative law*** **:**

Addition and multiplication of numbers are associative,

i.e., $a+(b+c)=(a+b)+c,\ a\times(b\times c)=(a\times b)\times c.$

Union and intersection of sets are also associative,

i.e., $A\cup(B\cup C)=(A\cup B)\cup C,$
$A\cap(B\cap C)=(A\cap B)\cap C.$

(iii) ***Distributive law*** **:**

In algebra of numbers only *one* law operates.

$$a\times(b+c)=a\times b+a\times c\,.$$
$$a+(b\times c)\neq(a+b)\times(a+c)$$

In algebra of sets, we have

$$A\cap(B\cup C)=(A\cap B)\cup(A\cap C)$$

(which is same as ordinary algebra shown above)

and $A\cup(B\cap C)=(A\cup B)\cap(A\cup C)$

(this additional law holds good only in algebra of sets)

(iv) ***Idempotent law*** **:**

This law also shows that the algebra of sets is not completely analogous to that of ordinary algebra of numbers.

In algebra of numbers we have,

$$a+a=2a \text{ and } a\times a=a^2.$$

But for the set A

$$A\cup A=A \text{ and } A\cap A=A.$$

(v) ***Identity law :***

In ordinary algebra 0 is taken as an identity element for addition only, *i.e.*, $a+0=a$. In algebra of sets, union of a null set ϕ with any set A is the set A itself, *i.e.*, $A\cup\phi=A$.

Again, in ordinary algebra, 1 is taken as an identity element for multiplication, since $a\times 1=1$. In algebra of sets, $A\cap U=A$, where U is Universal set.

For such similarities (as shown) $A\cap B$ is known as logical sum and $A\cap B$ as logical product.

(vi) ***Complement law :***

In algebra of numbers if a is a fraction, say $\frac{1}{3}$, then its complement is $\frac{2}{3}$, where

$$\tfrac{1}{3}+\tfrac{2}{3}=1,\ \tfrac{1}{3}\times\tfrac{2}{3}=\tfrac{2}{9}\ (\neq 0)$$

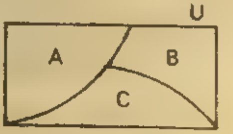
In algebra of sets, for every subset A of the universel set U, there is one and only one complement of A, *i.e.*, A′ such that $A\cup A'=U$, $A\cap A'=\phi$.

(vii) ***De Morgon's law :***

If A and B are two sets, then this law states that $(A\cup B)'=A'\cap B'$ and $(A\cap B)'=A'\cup B'$.

Partition of a Set.

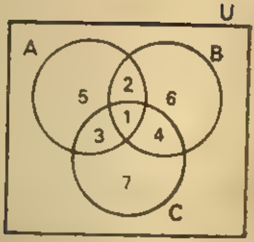
A universal set U may have a number of disjoint subsets. If again these subsets are joined together, then the same universal set is formed, as shown in the figure.



In the disjoint sets, no element is common. If, however, we take in account the elements which may be common to two or more subsets, then we will find more partitions. In case of *three subsets* of a universal set, there will be *eight partitions* (*i.e.*, $2^3=8$) as shown by symbols and Venn Diagram side by side.

By Symbols :

1. $A \cap B \cap C$
2. $A \cap B \cap C'$
3. $A \cap B' \cap C$
4. $A' \cap B \cap C$
5. $A \cap B' \cap C'$
6. $A' \cap B \cap C'$
7. $A' \cap B' \cap C$
8. $A' \cap B' \cap C'$

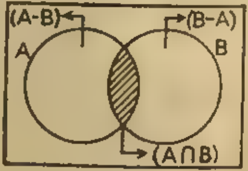


If again we take unions, intersections and complement of the three subsets, we get more regions as follows :

Subsets	*Regions*
$A \cup B \cup C$	1 to 7
$(A \cup B) \cap C$ or $(A \cap C) \cup (B \cap C)$	1, 3, 4
$A' \cup (A \cap C)$	1, 3, 4, 6, 7, 8
$(A \cap B) \cup C$	1, 2, 3, 4, 7
$(A \cup B) \cap C'$	2, 5, 6
$A' \cap (A \cap C)$	None

Number of Elements in a Set.

In a finite set, if operations are made, some new subsets will be formed. In this section we will find the values of these new subsets. Since A is a finite set, we shall denote it by $n(A)$ for the finite elements in A, which may be obtained by actual counting. But for unions of two or more sets, we have different formulas.



(1) FOR UNION OF TWO SETS :

For two sets A and B which are not disjoint,

$$n(A \cup B) = n(A) + n(B) - n(A \cup B)$$

Proof : From the Venn Diagram, we observe that $A \cup B$ is the union of three mutually disjoint sets $(A-B)$, $A \cap B$ and $(B-A)$.

$$\therefore \quad n(A \cup B) = n(A-B) + n(A \cap B) + n(B-A) \qquad (1)$$

Since A and B are finite sets, let us assume that $n(A) = x$, $n(B) = y$ and $n(A \cap B) = z$, then $n(A-B) = x - z$ and $n(B-A) = y - z$.

$\therefore$ from (1), we get

$$n(A \cup B) = x - z + z + y - z = x + y - z = n(A) + n(B) - n(A \cap B) \qquad (2)$$

If $A\cup B=\phi$ then $n(A\cup B)=n(A)+n(B)$,
i.e., if A and B are disjoint sets, then $n(A\cup B)=n(A)+n(B)$

(2) For Union of Three Sets :

Let A, B and C are the three sets (*not* mutually disjoint) then

$$\begin{aligned}
n(A\cup B\cup C) &= n[A\cup(B\cup C)]\\
&= n(A)+n(B\cup C)-n[A\cap(B\cup C)]\\
&= n(A)+n(B)+n(C)-n(B\cap C)-n[(A\cap B)\cup(A\cap C)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\text{(distributive law)}\\
&= n(A)+n(B)+n(C)-n(B\cap C)\\
&\qquad\qquad -[n(A\cap B)+n(A\cap C)-n(A\cap B\cap C)],\\
&\qquad\text{since } [(A\cap B)\cap(A\cap C)]=(A\cap B\cap C)\\
&= n(A)+n(B)+n(C)-n(A\cap B)-n(B\cap C)-n(C\cap A)\\
&\qquad\qquad +n(A\cap B\cap C).
\end{aligned}$$

If again A, B and C are *mutually disjoint*, then we find

$$n(A\cup B\cup C)=n(A)+n(B)+n(C).$$

***Splitting*:**

The basic set A, B or C may be split up into a number of intermediate sub-groups or sets, as follows :

$$\begin{aligned}
n(A) &= n(A\cap B)+n(A\cap B')\\
&= n(A\cap B\cap C)+n(A\cap B\cap C')+n(A\cap B'\cap C)+n(A\cap B'\cap C')
\end{aligned}$$

$n(A)$ can also be split up in the form of A, C.

Now the subsets of the union of three sets will be

$$\begin{aligned}
n(A\cup B\cup C) &= n(A\cap B'\cap C')+n(A\cap B\cap C')+n(B\cap A'\cap C')\\
&\quad +n(A\cap B'\cap C)+n(A\cap B\cap C)+n(A'\cap B\cap C)+n(A'\cap B'\cap C)\\
&\qquad\qquad [\text{ the residual subset is } n(A'\cap B'\cap C')\,]
\end{aligned}$$

***Re-grouping*:**

The subsets of a universal set may be re-grouped with other subset. Re-grouping or splitting (shown before) of subset will depend on the nature of the problem. Process of re-grouping is shown below (with reference to the above diagram of A, B and C sets) in few cases,

$$\begin{aligned}
n(A\cap C') &= n(A)-n(A\cap C)\\
n(B\cap A') &= n(B)-n(B\cap A)\\
n(C\cap B') &= n(C)-n(C\cap B)\\
n(A\cap B) &= n(A\cap B\cap C)+n(A\cap B\cap C')\\
n(B\cap C) &= n(A\cap B\cap C)+n(B\cap C\cap A')
\end{aligned}$$

or $n(B\cap C\cap A') = n(B\cap C) - n(A\cap B\cap C)$

$n(A\cap B'\cap C') = n(A) - n(A\cap B) - n(A\cap C) + n(A\cap B\cap C)$

$n(B\cap C'\cap A') = n(B) - n(B\cap C) - n(B\cap A) + n(A\cap B\cap C)$

$n(A) = n(A\cap B) + n(A\cap B')$

$n(C) = n(B\cap C) + n(C\cap B')$.

Worked out Examples :

1. In a class of 100 students, 45 students read Physics, 52 students read Chemistry and 17 students read both the subjects. Find the number of students who study neither Physics nor Chemistry.

We know $n(A\cup B) = n(A) + n(B) - n(A\cap B)$.

Here $n(A) = 45$, $n(B) = 52$, $n(A\cap B) = 17$

so, $n(A\cup B) = 45 + 52 - 17 = 97 - 17 = 80$

We are to find $n(A'\cap B')$, which is $100 - 80 = 20$.

2. In a class of 50 students, 15 read Physics, 20 read Chemistry and 20 read Mathematics. 3 read Physics and Chemistry, 6 read Chemistry and Mathematics, and 5 read Physics and Mathematics. 7 read none of the three subjects. How many students read all the three subjects.

Here, $n(A\cup B\cup C) = 50$, $n(A) = 15$, $n(B) = 20$, $n(C) = 20$

$n(A\cap B) = 3$, $n(B\cap C) = 6$, $n(C\cap A) = 5$, $n(A\cap B\cap C) = ?$

Using the formula,

$$n(A\cup B\cup C) = n(A) + n(B) + n(C) - n(A\cap B) - n(B\cap C) - n(C\cap A) + n(A\cap B\cap C)$$

or, $50 = 15 + 20 + 20 - 3 - 6 - 5 + n(A\cap B\cap C)$

or, $n(A\cap B\cap C) = 9$. But 7 students read nothing.

$\therefore$ required no. of students $= 9 - 7 = 2$.

(By re-grouping of sets)

3. An enquiry into 1000 candidates who failed in C.A. final examination revealed the following data :—

658 failed in aggregate, 166 failed in Aggr. and Group I
372 „ „ Group I, 434 „ „ „ „ Group II
590 „ „ „ II, 126 „ „ „ both Groups

Find how many candidates failed in :

(*a*) All the three,

(*b*) In aggregate but not in Group II,

(*c*) Group I but not in aggregate,

(*d*) Group II but not in aggregate,

(*e*) Aggregate but not in Group I and Group II.

Let A, B and C are the sets of students who have failed in aggregate, group I and group II respectively.

(*a*) $n(A \cup B \cup C) = n(A) + n(B) + n(C) - n(A \cap B) - n(B \cap C) - n(A \cap C) + n(A \cap B \cap C)$

$\Rightarrow 1000 = 658 + 372 + 590 - 166 - 434 - 126 + n(A \cap B \cap C)$

or, $n(A \cap B \cap C) = 106.$

(*b*) Failed in Aggr. but not in Gr. II $= n(A \cap C')$

$n(A \cap C') = n(A) - n(A \cap C)$

[as $A \cap C'$ and $A \cap C$ are disjoined sets]

$= 658 - 434 = 224.$

(*c*) Failed in Gr. I but not in Aggr. $= n(B \cap A')$

$n(B \cap A') = n(B) - n(B \cap A)$

[as $B \cap A'$ and $B \cap A$ are disjoined sets]

$= 372 - 166 = 206.$

(*d*) Failed in Gr. II but not in Aggr. $= n(C \cap B')$

$n(C \cap B') = n(C) - n(C \cap B)$

$= 590 - 126 = 464.$

(*e*) Failed in Aggr. but not in Gr. I and Gr. II

$n(A \cap B' \cap C') = n(A) - n(A \cap B) - n(A \cap C) + n(A \cap B \cap C)$

$= 658 - 166 - 434 + 106 \cdots$ from (*a*)

$= 164.$

USE OF VENN DIAGRAM :

4. In a survey of 150 students, it was found that 40 students studied Physics, 60 students studied Chemistry and 50 students studied Mathematics, and 15 students studied all the three subjects. 27 students studied Physics and Chemistry, 35 students studied Chemistry and Mathematics and 25 students Physics and Mathematics. Find the number who studied only Physics and the number who studied none of these subjects.

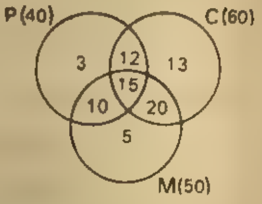
Let P, C and M represent Physics, Chemistry and Mathematics respectively.

Now, $P \cap C \cap M = 15$

$P \cap C = 27$

$C \cap M = 35$

$P \cap M = 25$



Students studying Physics only

$= 40 - (12 + 15 + 10) = 3$

(see the diagram)

Number of students who studied one or more subjects

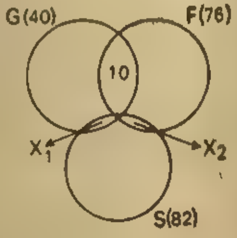
$= 15 + 12 + 10 + 20 + 3 + 13 + 5 = 78$

$\therefore$ number of students who studied none

$= 150 - 78 = 72.$

5. Out of a certain set of 200 students, 40 read German, 76 read French and 82 read Spanish, 36 read exactly two of these languages, but none read all the three. 34 read German but not Spanish and 10 read both German and French. Find how many of the original set fail to read any of the three languages, and how many students read German only.

Total number of students reading one or more of the subjects $= 40 + 76 + 82 = 198$.



Here, there is no student reading all the three languages together, so there will be no common region between the three circles, which means three circles will meet at a point.

Now 36 students read exactly two subjects, *i.e.*, German and French, French and Spanish, German and Spanish.

Again 10 students read German and French.

So, $x_1 + x_2 = 36 - 10 = 26.$

Again 34 students read German but not Spanish.

So, $x_1 = 40 - 34 = 6.$ $\therefore$ $x_2 = 26 - 6 = 20.$

Number of students read German only $= 40 - 10 - 6 = 24.$

Total number of students reading German only, French only, Spanish only $= 198 - 36 = 162.$

$\therefore$ students fail to read either of the three languages

$= 200 - 162 = 38.$

Cartesian Product.

Let A and B be two given sets. If $a \in$ A, $b \in$ B then (a, b) denotes an ordered pair, a is regarded as the first element and b the second element so that (a, b) is *not the same as* (b, a). In case of a set $\{a, b\} = \{b, a\}$.

Two ordered pairs (a, b) and (c, d) will be equal if and only if $a = c$ and $b = d$,

i.e., $(a, b) = (c, d) \Rightarrow a = c, b = d$.

Definition :

If A and B are two sets, then the set of all ordered pairs (a, b) such that $a \in$ A, $b \in$ B is called the cartesian product of A and B and is denoted by A × B.

In symbols :

$$A \times B = \{x : x = (a, b), a \in A, b \in B\}$$

EXAMPLE :

$$A = \{1, 2, 3\}, \ B = \{1, 2\}$$
$$A \times B = \{(1, 1), (1, 2), (2, 1), (2, 2), (3, 1), (3, 2)\}$$

EXAMPLE :

If $A = \{1, 2, 3\}$, $B = \{2, 3\}$

Prove that, $A \times B \neq B \times A$. [C. A. Entr., May '74]

$$A \times B = \{(1, 2)(1, 3), (2, 2)(2, 3) (3, 2)(3, 3)\}$$
$$B \times A = \{(2, 1)(2, 2)(2, 3)(3, 1)(3, 2)(3, 3)\}.$$

We find the elements (1, 2)(1, 3) of A × B, are not the elements of B × A.

$$\therefore \ A \times B \neq B \times A.$$

Properties :

1. Since two ordered pairs (a, b), (b, a) are unequal the cartesian product is not commutative

   i.e., $A \times B \neq B \times A$ unless $A = B$ or one set is empty.

2. If set A has m elements, and set B has n elements then set A × B has mn elements.

3. A × B is empty if either A or B is empty.

4. A × B is infinite if either A or B is infinite and the other is non-empty.

Cartesian Product of n Sets :

Let $A_1, A_2 \cdots\cdots A_n$ be n sets. The set of ordered n-tuples $(a_1, a_2, \cdots\cdots a_n)$ where $a_i \varepsilon A_i$, $i=1, 2, \cdots\cdots n$ is known as cartesian product of $A_1, A_2, \cdots\cdots, A_n$ and is denoted by $A_1 \times A_2 \times \cdots\cdots \times A_n$.

EXAMPLE :

If $A=\{2, 3\}$, $B=\{1, 3\}$, $C=\{3, 4\}$.

Find (i) $A\times(B\cap C)$ (ii) $A\times(B\cap C)$ (iii) $(A\times B)\cup(B\times C)$.

(i) $A\times(B\cup C)=\{2, 3\}\times\{1, 3\} \cup \{3, 4\}$

$=\{2, 3\}\times\{1, 3, 4\}$

$=\{(2, 1), (2, 3), (2, 4), (3, 1), (3, 3), (3, 4)\}$.

(ii) $A\times(B\cap C)=\{2, 3\}\times\{1, 3\}\cap\{3, 4\}$

$=\{2, 3\}\times\{3\}=\{(2, 3)(3, 3)\}$.

(iii) $A\times B=\{(2, 1)(2, 3), (3, 1)(3, 3)\}$

$B\times C=\{(1, 3)(1, 4), (3, 3)(3, 4)\}$

$\therefore$ $(A\times B)\cup(B\times C)=\{(2, 1)(2, 3)(3, (13), 3)(1, 3)(1, 4)(3, 4)\}$.

Useful Results.

1. Prove that (i) $A\subset B$ and $C\subset D \Rightarrow (A\times C)\subset(B\times D)$

(ii) $A\times(B\cup C)=(A\times B)\cup(A\times C)$

[C. A. Entr., Nov. '74]

(iii) $A\times(B\cap C)=(A\times B)\cap(A\times C)$

(iv) $(A\times B)\cap(S\times T)=(A\cap S)\times(B\cap T)$

(i) Let (a, c) be any element in $(A\times C)$.

then $(a, c)\ \varepsilon\ (A\times C) \Rightarrow a\ \varepsilon\ A$ and $c\ \varepsilon\ C$

$\Rightarrow a\ \varepsilon\ B$ and $c\ \varepsilon\ D$, as $A\subset B$, $C\subset D$

$\Rightarrow (a, c)\ \varepsilon\ (B\times D)$

$\therefore$ $(A\times C)\subset(B\times D)$.

(ii) Let $(x, y)\ \varepsilon\ A\times(B\cup C)$

$\Rightarrow x\ \varepsilon\ A$ and $y\ \varepsilon\ (B\cup C)$

$\Rightarrow x\ \varepsilon\ A$ and $(y\ \varepsilon\ B$ or $y\ \varepsilon\ C)$

$\Rightarrow (x\ \varepsilon\ A$ and $y\ \varepsilon\ B)$ or $(x\ \varepsilon\ A$ and $y\ \varepsilon\ C)$

$\Rightarrow (x, y)\ \varepsilon\ (A\times B)$ or $(x, y)\ \varepsilon\ (A\times C)$

$\Rightarrow (x, y)\ \varepsilon\ (A\times B)\cup(A\times C)$

$\therefore$ $A\times(B\cup C)\subseteq(A\times B)\cup(A\times C)$.

Again, let $(u, v) \varepsilon (A \times B) \cup (A \times C)$

$\Rightarrow (u, v) \varepsilon (A \times B)$ or $(u, v) \varepsilon (A \times C)$

$\Rightarrow (u \varepsilon A$ and $v \varepsilon B)$ or $(u \varepsilon A$ and $v \varepsilon C)$

$\Rightarrow u \varepsilon A$ and $(v \varepsilon B$ or $v \varepsilon C)$

$\Rightarrow (u, v) \varepsilon A \times (B \cup C)$

$\therefore \quad (A \times B) \cup (A \times C) \subseteq A \times (B \cup C)$

$\therefore \quad A \times (B \cup C) = (A \times B) \cup (A \times C)$

(iii) Similar to above, and it is left to the students.

(iv) Let $(x, y) \varepsilon (A \times B) \cap (S \times T)$

$\Rightarrow (x, y) \varepsilon (A \times B)$ and $(x, y) \varepsilon (S \times T)$

$\Rightarrow (x \varepsilon A$ and $y \varepsilon B)$ and $(x \varepsilon S$ and $y \varepsilon T)$

$\Rightarrow (x \varepsilon A$ and $x \varepsilon S)$ and $(y \varepsilon B$ and $y \varepsilon T)$

$\Rightarrow x \varepsilon (A \cap S)$ and $y \varepsilon (B \cap T)$

$\Rightarrow (x, y) \varepsilon (A \cap S) \times (B \cap T)$

$\therefore \quad (A \times B) \cap (S \times T) \subseteq (A \cap S) \times (B \cap T)$

Again, let $(u, v) \varepsilon (A \cap S) \times (B \cap T)$

$\Rightarrow u \varepsilon (A \cap S)$ and $v \varepsilon (B \cap T)$

$\Rightarrow (u \varepsilon A$ and $u \varepsilon S)$ and $(v \varepsilon B$ and $v \varepsilon T)$

$\Rightarrow (u, v) \varepsilon (A \times B)$ and $(u, v) \varepsilon (S \times T)$

$\Rightarrow (u, v) \varepsilon (A \times B) \cap (S \times T)$

$\therefore \quad (A \cap S) \times (B \cap T) \subseteq (A \times B) \cap (S \times T)$

$\therefore \quad (A \times B) \cap (S \times T) = (A \cap S) \times (B \cap T).$

2. If $A \subset B$, show $(A \times A) \Rightarrow (A \times B) \cap (B \times A)$

Let $(x, y) \varepsilon (A \times A)$

then $x \varepsilon A$ and $y \varepsilon A$. Now as $A \subset B$ by hypothesis

$x \varepsilon A \Rightarrow x \varepsilon B$ and $y \varepsilon A \Rightarrow \bar{y} \varepsilon B$

$\therefore \quad (x, y) \varepsilon (A \times A) \Rightarrow x \varepsilon A$ and $y \varepsilon A$

$\Rightarrow x \varepsilon A$ and $y \varepsilon B$

$\Rightarrow (x, y) \varepsilon (A \times B).$

Again, $(x, y) \varepsilon (A \times A) \Rightarrow x \varepsilon A$ and $y \varepsilon A$

$\Rightarrow x \varepsilon B$ and $y \varepsilon A$

$\Rightarrow (x, y) \varepsilon (B \times A).$

Thus $(x, y)\ \varepsilon\ (A \times A) \Rightarrow (x, y)\ \varepsilon\ (A \times B)$ also $\varepsilon\ (B \times A)$

$$\Rightarrow (x, y)\ \varepsilon\ (A \times B) \cap (B \times A)$$

$$\therefore\quad (A \times A) \Rightarrow (A \times B) \cap (B \times A)$$

3. For any three sets A, B and C, show that

$$A \times (B \sim C) = (A \times B) \sim (A \times C)$$

Let $(x, y)\ \varepsilon\ A \times (B \sim C) \Rightarrow x\ \varepsilon\ A$ and $y\ \varepsilon\ (B \sim C)$

$\Rightarrow x\ \varepsilon\ A$ and ($y \in B$ and $y \notin C$)

$\Rightarrow$ ($x\ \varepsilon\ A$ and $y\ \varepsilon\ B$) and ($x \in A$ and $y \notin C$)

$\Rightarrow (x, y) \in (A \times B)$ but $(x, y) \notin (A \times C)$

$\Rightarrow (x, y)\ \varepsilon\ (A \times B) \sim (A \times C)$

$\therefore\quad A \times (B \sim C) \subseteq (A \times B) \sim (A \times C).$

Again, let $(u, v)\ \varepsilon\ (A \times B) \sim (A \times C)$

$\Rightarrow (u, v) \in (A \times B)$ but $(u, v) \notin (A \times C)$

$\Rightarrow$ ($u\ \varepsilon\ A$ and $v\ \varepsilon\ B$) and ($u \in A$ and $v \notin C$)

$\Rightarrow u\ \varepsilon\ A$ and ($v \in B$ and $v \notin C$)

$\Rightarrow u\ \varepsilon\ A$ and $v\ \varepsilon\ (B \sim C)$

$\Rightarrow (u, u)\ \varepsilon\ A \times (B \sim C)$

$\therefore\quad (A \times B) \sim (A \times C) \subseteq A \times (B \sim C)$

Thus, $A \times (B \sim C) = (A \times B) \sim (A \times C).$

EXERCISE 13

1. Given $A = \{1, 2, 3, 4, 5,\}$ $B = \{2, 4, 6\}$
   $C = \{3\}$ $D = \{0, 1, 2, \cdots\cdots 9\}.$

Find (i) $A \cup C$, (ii) $A \cup (B \cup C)$, (iii) $B \cap C$, (iv) $C \cap D$, (v) $A \cap (B \cap C)$, (vi) $(A \cap B) \cap C$, (vii) $A \triangle B$.

2. Given U (universal) $= \{0, 1, \cdots\cdots 9)$, $A = \{2, 4, 6\}$,
   $B = \{1, 3, 5, 7),$ $C = \{6, 7\}.$

Find (i) $A' \cap B$, (ii) $(A \cup B) \sim C$, (iii) $(A \cup C)'$, (iv) $(A \cap U) \cap (B \cup C)$.

3. If S be the set of all prime numbers, $M = \{0, 1, 2, \cdots 9\}$, exhibit (i) $S \cap M$, (ii) $M - (S \cap M)$.

4. Let $A = \{a, b, c\}$, $B = \{d\}$, $C = \{c, d\}$,
   $D = \{a, b, d\}$, $E = \{a, b\}$.

Determine if the following statements are true ?

(i) $E \subset A$, (ii) $B \subset C$, (iii) $A \subset D$, (iv) $C \subset D$,
(v) $E = C$, (vi) $B \supset C$, (vii) $A \sim D$.

5. Determine which of the following sets are same
$A = \{5, 7, 6\}$, $B = \{6, 8, 7\}$, $C = \{5, 6, 7\}$
$D = \{x : x$ is an integer greater than 2 but less than $6\}$
$E = \{1, 2, 3, 4, 5, 6\}$ $F = \{3, 4, 5\}$.

6. Fill up the blanks by appropriate symbol
$\in, \notin, \subset, \subseteq, \supset$
(i) $3 : \cdots\cdots \{3, 4\} \cup \{4, 5, 6\}$
(ii) $\{6\} \cdots\cdots \{5, 6\} \cap \{6, 7, 8\}$
(iii) $\{3, 4, 5\} \cdots\cdots \{2, 3, 4\} \cup \{3, 4, 5\}$
(iv) $\{a, b\} \cdots\cdots \{a\}$
(v) $4 \cdots\cdots \{3, 5\} \cup \{5, 6, 7\}$
(iv) $\{1, 2, 2, 3\} \cdots\cdots \{3, 2, 1\}$.

7. Find the power set P (A) of the set $A = \{a, b, c\}$.

8. Indicate which of the sets is a null set ?
$X = \{x : x^2 = 4, 3x = 12\}$
$Y = \{x : x + 7 = 7\}$
$Z = \{x : x \neq x\}$.

9. If $U = \{x : x$ is a letter in English alphabet$\}$
$V = \{x : x$ is a vowel$\}$
$W = \{x : x$ is a consonant$\}$
$Y = \{x : x$ is e or any letter before e in the alphabet$\}$
$Z = \{x : x$ is e or any one of the next four letters$\}$

Find each of the following sets :

(i) $U \cap V$, (ii) $V \cap Y$, (iii) $Y \cap Z$,
(iv) $U \cap W'$, (v) $U \cap (W \cap V)$.

10. Given $A = \{x : x \,\varepsilon\, N$ and x is divisible by $2\}$
$B = \{x : x \,\varepsilon\, N$ and x is divisible by $3\}$
$C = \{x : x \,\varepsilon\, N$ and x is divisible by $4\}$
Describe $A \cap (B \cap C)$.

11. If $A = \{x : x \,\varepsilon\, N$ and $x \leqslant 6\}$
$B = \{x : x \,\varepsilon\, N$ and $3 \leqslant x \leqslant 8\}$
$U = \{x : x \,\varepsilon\, N$ and $x > 10\}$

Find the elements of the following sets with any remark of any :

(i) $(A \cup B)'$, (ii) $A' \cap B'$, (iii) $(A \cap B)'$, (iv) $A' \cup B'$.

12. If S be any set, P (S) its power set and if A and B belong to P (S), then show that $B \cap (A \sim B) = \phi$.

13. For any sets A and B, show that

(i) $A \sim B = A \sim (A \cap B)$

(ii) $A' \sim B' = B \sim A$.

14. For any three sets A, B and C, show that

$$A \cup (B \sim C) \neq (A \cup B) \sim (A \cup C)$$

15. Let $A = (a, b)$, $B = (b, c)$, $C = (d, e)$.

Find (i) $A \times (B \cup C)$

(ii) $(A \cap B) \times C$

(iii) $(A \times B) \cap (A \times C)$

16. If $A = \{1, 4\}$, $B = \{2, 3\}$, $C = \{3, 5\}$

Prove that $A \times B \neq B \times A$ [C. A. Entr., May 75]

Also find $(A \times B) \cap (A \times C)$.

17. If $A = \{1, 2, 3\}$, $B = \{2, 3, 4\}$, $S = \{1, 3, 4\}$, $T = \{2, 4, 5\}$,

verify that $(A \times B) \cap (S \times T) = (A \cap S) \times (B \cap T)$

[C. A. Entr., Nov. 76]

18. In a class of 30 students, 15 students have taken English, 10 students have taken English but not French. Find the number of students who have taken (i) French, and (ii) French but not English.

(Ans. : 20, 15)

19. In a survey of 320 persons, number of persons taking tea is 210, taking milk 100 and coffee is 70. Number of persons who take tea and milk is 50, milk and coffee is 30 and tea and coffee is 50. The number of persons taking all the three together is 20. Find the number of persons who take neither tea, nor milk nor coffee.

(Ans. : 50)

20. Out of 440 boys in a College, 112 boys read German, 120 read French and 168 Spanish. Of these 32 read French and Spanish, 40 read German and Spanish, 20 read German and French, while 12 read all the three languages. How many boys

(i) did not read any language,

(ii) read just one language ? (Ans. : 320, 252)

21. In a survey of 100 students, the number of students studying various languages is as follows :

German only 18, German but not Spanish 23, German and French 8, German 26, French 48, French and Spanish 8, no language 24.

Find (i) how many students took Spanish ?

(ii) how many took German and Spanish but not French ?
(Ans. : 18, 0)

22. A reporter supplied the following data about another set of 100 boys—

All three languages 5, German and Spanish 10, French and Spanish 8, German and French 20, Spanish 30, German 23, French 50.

The reporter was dismissed, why ? (Ans. : Negative number)

23. In a survey concerning the smoking habits of consumers it was found that 55% smoke cigarette A, 50% smoke B, 42% smoke C, 28% smoke A and B, 20% smoke A and C, 12% smoke B and C and 10% smoke all the three cigarettes.

(i) What % age do not smoke ?

(ii) What % age smoke exactly 2 brands of cigarettes ?
(Ans. : 3%, 30%)

24. A company studies the product preferences of 10,000 consumers. It was found each of the products A, B, C was liked by 5000, 3470 and 4830 respectively and all the products were liked by 500 ; products A and B were liked by 1000, products B and C were liked by 900 and products C and A were liked by 1400. Prove that the study results are not correct. It was found that an error was made in recording the number of consumers liking the products B and C. What is the value of this number ? (Ans. : 1400).

25. In a survey of 1000 boys playing outdoor games of football, cricket and hockey were recorded. Each boy plays any one of the games, 400 boys did not play hockey, 370 did not play cricket and 550 did not touch football. 300 played football and cricket, 270 played both cricket and hockey, 200 both hockey and football. How many boys played all the three games ? How many played only football ?
(Ans. : 90)

26. In a close quarter battle 60% of combatants lost an eye, 75% lost an arm and 80% lost a leg. If any of the combatants has at least one of these types of loss and the percentage of combatants having loss of exclusively any two types at the same time is 65. Calculate the percentage of combatants having the three types of loss at the same time. (Ans. : 25%)

14

PROBABILITY

Introduction and Meaning

In nature or in our day-to-day life, situations arise in which we can not predict with absolute certainty about the exact occurrence of any future event. We may hope, speculate or guess with or without reason about the happening of an event only. Now the likelihood of the occurrence is expressed by the term *Probability*. And a distinct branch of Mathematics has gradually developed since the last sixteenth century formulating different theories on probabilities.

Initially the applications of probability theories were restricted to games of chances. But in course of time they are being incorporated in the business processes and decision making apparatus by business firms, governments, and professional and non-profit organizations.

Predictions on demand for a new product, estimations of production costs, forecasting crop failures, buying insurance, preperation of a budget, etc.—all are better enumerated with the help of the Mathematics of probability, as they have some element of chance inherent in them. And the advantage of probability theory lies in the fact that it has the ability to quantify "how likely" some event is. We have three different ways of calculating or cumputing probabilities. There are two objective methods and one subjective method. The objective ones relate to the classical approach and the empirical approach respectively.

The classical approach, *which forms the subject-matter of this chapter*, is generally made when the given situations have equally likely outcomes. Games of chance, which often involve coin-tossing, rolling dice, or drawing cards, usually have this characteristic of equally likely outcomes.

The empirical approach is based on the relative frequency of occurrence of an event over a large number of repeated trials. When this approach is made, the following important points should be noted :

1. The probability so determined is only an estimate of the actual value.
2. The larger the number of trials, the better the estimate of the probability.

3. The trials should be conducted under identical conditions.

The subjective method is based on an individual's personal feeling such as the judgement that 'there is 90% chance that it will be rain tomorrow' or that 'there is a better than 50% chance that a labour strike will be settled this time'.

We will not, however, make use of this method in this text.

Random Experiment

A random experiment is an experiment whose all possible results (outcomes) are known and which can be repeated under identical conditions but it is not possible to predict the outcome of any particular trial in advance.

Note. Any particular performance of a random experiment is called trial.

Tossing of a coin is a noteworthy example of a random experiment, the outcomes are two in number *a head* or *a tail* appears but exact prediction is not possible in any tossing.

Similarly, throwing of a die is also a random experiment with six outcomes—either 1, 2, 3, 4, 5, or 6 will turn up, but here as before, exact prediction is impossible in any throwing.

Any process of observation in business, as for example, the production of a commodity on different days or price of a commodity in different months may be taken as outcomes of a random experiment.

Events

The possible outcomes of a random experiment are called *events.*

Elementary and Compound (Composite) events :

Elementary (simple) event is an outcome of an experiment that cannot be decomposed further, whereas *Compound (composite) event* is an aggregate of some *elementary* or *simple* events and is decomposable into *simple* events.

In the experiment of tossing a coin, one simple event is 'head' and the other is 'tail' but the event 'head or tail' is a compound (composite) event since it can be decomposed into two simple events the event 'head' and the event 'tail'. In tossing two coins, the event 'both heads' (HH) is a simple event and so also the event 'both tails' (TT); but the event 'one head and one tail' is a compound event consisting of two events (HT) and (TH). [H is for head and T for tail.]

In throwing two dice the event 'total 12 points', viz. (6, 6) is a simple event but the event 'total 7 points' is a Compound (composite) event consisting of 6 simple events (1, 6), (2, 5), (3, 4), (4, 3), (5, 2) and (6, 1).

Mutually exclusive events :

Two events are said to be *mutually exclusive* or *incompatible* when the occurrence of one of them excludes the occurrence of the other or in other words, two events are mutually exclusive if both cannot occur simultaneously.

If a single coin is tossed, the 'head' and the 'tail' cannot occur in the same trial. Hence the event 'head' and the event 'tail' are mutually exclusive. If two coins are tossed, the events (HH), (HT), (TH) and (TT) are mutually exclusive.

In drawing a single card from a pack of well-shuffled 52 cards the events 'card is a spade' and 'card is a club' are mutually exclusive, because a card cannot both be a spade and a club. But the events 'card is a spade' and 'card is a face card' are not mutually exclusive, since some spade cards are face cards.

Exhaustive events :

Events are said to be *exhaustive* if at least one of them must necessarily occur.

The total number of all possible outcomes of a random experiment will constitute an exhaustive set of events.

Thus, in tossing of a coin there are two exhaustive events 'head' and 'tail' and in throwing of a die the exhaustive events are six—either 1, 2, 3, 4, 5 or 6. In drawing a single card from a pack of 52 cards the events 'card is red' and 'card is black' are collectively exhaustive.

Equally likely events :

Events are said to be *equally likely* if after taking into account all relevant evidences, no one of the events can be expected to occur in preference to the other events, that is, when one does not occur more often than the other.

In tossing of a coin, 'head' and 'tail' are equally likely events. All the six faces of a die are equally likely events, when it is thrown. All the 52 cards of an well-shuffled pack of cards are equally likely events when one card is drawn.

Certain and Impossible events :

An event is called *certain* or *sure* when all the possible outcomes of an experiment are favourable to the event, whereas an event is called *impossible* when none of the outcomes is favourable to the event.

Classical definition of Probability.

This concept of probability happens to be the most primitive one and depends upon the notion of equally likely events. If for a random experiment there is n (finite) mutually exclusive, exhaustive and equally likely outcomes and r of them are favourable to an event A, then the probability of the event A is defined and denoted by

$$P(A) = \frac{r}{n}.$$

Thus, when the possible outcomes are equally likely, then the ratio of the number of ways an event A can occur to the total number of possible outcomes is the probability of the *event* A.

Now, by definition, $P(A) = \frac{r}{n}$ must always lie between 0 and 1, since $0 \leq r \leq n$. When $r = n$, *i.e.*, when event is *certain* and $P(A) = 1$ and when $r = 0$, *i.e.*, when the event is *impossible* and $P(A) = 0$. So, $0 \leq P(A) \leq 1$.

For the probability of the event 'the sun will not rise tomorrow' is zero. The event is an *impossible* event. And, the probability of the events 'man will die some day' is 1. This is a *certain* event.

Remark. If the event not-A is denoted by $\overline{A}$, then from the definition of probability it follows that, $P(\overline{A}) = \frac{n-r}{n} = 1 - \frac{r}{n} = 1 - P(A)$.

Odds.

The ratio of probabilities for A and for $\overline{A}$ is often called *odds in favour* of the event A.

$$\therefore \quad \text{Odds in favour of the event } A = \frac{P(\overline{A})}{P(A)} = \frac{r/n}{(n-r)/n} = \frac{r}{n-r}.$$

And, the ratio of probabilities for $\overline{A}$ and for A is called *odds against* the event A.

$$\therefore \quad \text{Odds against the event } A = \frac{P(\bar{A})}{P(A)} = \frac{(n-r)/n}{r/n} = \frac{n-r}{r}.$$

EXAMPLE :

In a single toss of a fair coin, find the probability of getting 'head'.

SOLUTION :

There are two exhaustive, mutually exclusive and equally likely outcomes of the loss of a fair coin. Out of these *two* outcomes, only *one* is favourable to the event 'head'.

Hence, probability of getting a head $= \frac{1}{2}$.

EXAMPLE :

In a single toss of two fair coins, find the probability of obtaining (i) both heads, (ii) one head and one tail, and (iii) at least one tail.

SOLUTION :

There are *four* exhaustive, mutually exclusive and equally likely outcomes in a single toss of two fair coins, and they are (HH), (HT), (TH) and (TT) where H denotes 'head' and T denotes 'tail', of which only *one* outcome, viz. (HH) is favourable to the event 'both heads', *two* outcomes, viz. (HT) and (TH) to the event 'one head and one tail' and *three* outcomes (HT), (TH) and (TT) to the event 'at least one tail'.

Hence,

(i) P (both heads) $= \frac{1}{4}$

(ii) P (one head and one tail) $= \frac{2}{4} = \frac{1}{2}$

(iii) P (at least one tail) $= \frac{3}{4}$

EXAMPLE :

What is the probability of obtaining (i) an even number, (ii) a number less than 5, and (iii) 'a five' in a single throw of an unbiased die.

SOLUTION :

There are *six* mutually exclusive, exhaustive and equally likely outcomes, viz the appearance of the numbers 1, 2, 3, 4, 5 or 6. Of these outcomes *three* are favourable to the event 'an even number', viz. the appearance of 2, 4 or 6, *four* are favourable to the event 'number less than 5', viz. 1, 2, 3 or 4 and *one* is favourable to the event 'a five', viz. 5.

Hence,

(i) P (even number) $= \frac{3}{6} = \frac{1}{2}$

(ii) P (less than 5) $= \frac{4}{6} = \frac{2}{3}$

(iii) P (a five) $= \frac{1}{6}$.

EXAMPLE :

In a throw of two unbiased dice find the probability of throwing (i) total seven points, and (ii) total eight points.

SOLUTION :

In a throw of one unbiased die there are *six* exhaustive, mutually exclusive, and equally likely outcomes. So, when two dice are thrown, there are $6 \times 6 = 36$ exhaustive, mutually exclusive and equally likely outcomes. Of these 36 outcomes only in *six* cases, viz. (1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1) are favourable to the event 'total seven points'. Hence,

$$P\,(\text{total seven points}) = \tfrac{6}{36} = \tfrac{1}{6}$$

For the 'total eight points' the favourable cases are five, viz. (2, 6), (3, 5), (4, 4), (5, 3), (6, 2). Hence,

$$P\,(\text{total eight points}) = \tfrac{5}{36}.$$

EXAMPLE :

What is the probability of getting 3 white balls in a draw of 3 balls from a box containing 5 white and 4 black balls ? (C.A. 1976)

SOLUTION :

Total number of balls in the box $= 5 + 4 = 9$.

3 balls can be drawn from 9 balls in $^9C_3 = \dfrac{9 \times 8 \times 7}{3 \times 2 \times 1} = 84$ ways.

$\therefore$ Total number of possible cases $= 84$.

Total number of white balls in the box $= 5$.

3 white balls can be drawn from 5 white balls in

$$^5C_3 = \frac{5 \times 4 \times 3}{3 \times 2 \times 1} = 10 \text{ ways.}$$

Number of cases favourable to the event of getting 3 white balls $= 10$.

$\therefore$ Probability of drawing 3 white balls $= \dfrac{10}{84} = \dfrac{5}{42}$.

EXAMPLE :

A card is drawn at random from an well-shuffled pack of 52 cards. Find the probability and the odds that the card is a face-card (*i.e.*, king, queen or jack).

SOLUTION :

Total number of cards is 52. One can be drawn from 52 cards in 52 ways. So the total number of cutccmes is 52. There are 3 face-cards in each of 4 suits—a total of 12 face-cards, and one face-card from 12 face-cards can be drawn in 12 ways. Since the cards are well-shuffled and one card is drawn at random, each of the 52 cards is assumed equally likely to appear. Hence,

$$\text{P (face-card)} = \frac{12}{52} = \frac{3}{13}$$

$$\text{Odds in favour of a face-card} = \frac{12}{52-12} = \frac{12}{40} = \frac{3}{10}$$

$$\text{and, odds against a face-card} = \frac{52-12}{12} = \frac{40}{12} = \frac{10}{3}.$$

Limitations of Classical Definition :

(1) This definition is applicable when each outcome is equally likely. Being based on the idea of equally likely outcomes which means equally probable outcomes, the definition involves circular reasoning as the idea of probability has been used as a part of the definition of probability. This makes the definition unsatisfactory.

(2) It is not directly applicable when the total number of possible outcomes is infinite and also when it is not possible to enumerate all the possible outcomes.

(3) This definition is not applicable when the outcomes are not equally likely.

To remove these difficulties a second definition has been suggested as follows :

Statistical Definition : (Empirical Approach)

If an event A is found to occur r times when a random experiment is repeated n times, then r is called the frequency and $\frac{r}{n}$ is called the relative frequency of A. The limiting value of this relative frequency $\frac{r}{n}$ when n increases indefinitely is regarded as the probability of A connected with the experiment. Mathematically,

$$\text{P (A)} = \lim_{n \to \infty} \frac{r}{n}.$$

Thus the probability of an event A is the limit of the relative frequency of A in an infinite sequence of trials.

The exact determination of the probability of any event is not practically possible here since we are not sure that a limit to the relative frequency exists. On the other hand, we may use the relative frequency as an estimate of the probability of the event occurring under identical conditions. Hence, we have the following definition of the probability of the event A,

$$\text{P (A)} = \frac{\text{Number of times A occurred}}{\text{Total number of trials}}.$$

The following points are to be recognised when statistical definition is used :

(i) Probability determined here is only an estimate of the actual value.

(ii) The better the estimate of the probability, the larger the number of trials, and

(iii) The trials should be conducted under identical conditions.

Set Theoretic Approach.

To understand the probability theory and its potential for practical application, it is helpful to understand the basic principles of Set Theory.

Set

A set is a collection of items or objects having some common *characteristic* or *characteristics*.

Sample Space :

The set of all possible outcomes of an experiment is called the *sample space* of the experiment and is denoted by S. Each outcome of the experiment is called *an element* or a *sample point* of the sample space. The sample space is also called *Universal Set* or *Event Space* or *Possibility Set.*

A Sample Point is also called an *Event Point.*

Any outcome of the experiment corresponds to *exactly* one element in the sample space of the experiment.

A sample space may be *finite* or *infinite* according as it contains a finite or infinite number of sample points.

EXAMPLES :

(1) A random experiment of throwing a balanced coin has two outcomes *Head* and *Tail.* So, the sample space associated with this experiment consists of *two* sample points and may be expressed as :

S = {H, T} ; where H denotes head and T denotes tail of the coin.

(2) Sample space associated with the experiment of tossing a balanced coin twice or thrice are respectively,

S_1 = {HH, HT, TH, TT}

S_2 = {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}

S_1 consists of *four* and S_2 consists of *eight* sample points.

(3) Sample space associated with the experiment of throwing one dice consists of *six* sample points and is noted as :

$$S = \{1, 2, 3, 4, 5, 6\}.$$

(4) Sample space associated with the experiment of throwing two dice consists of 36 sample points and is expressed as :

S = {(1, 1) ; (1, 2) ; (1, 3) ; (1, 4) ; (1, 5) ; (1, 6) ; (2, 1) ; (2, 2) ;
(2, 3) ; (2, 4) ; (2, 5) ; (2, 6) ; (3, 1) ; (3, 2) ; (3, 3) ; (3, 4) ;
(3, 5) ; (3, 6) ; (4, 1) ; (4, 2) ; (4, 3) ; (4, 4) ; (4, 5) ; (4, 6) ;

(5, 1) ; (5, 2) ; (5, 3) ; (5, 4) ; (5, 5) ; (5, 6) ; (6, 1) ; (6, 2) ; (6, 3) ; (6, 4) ; (6, 5) ; (6, 6)}

(5) Sample space associated with the experiment of tossing a coin 10 times and recording the number of heads obtained is expressed as :

$$S=\{0, 1, 2, \cdots\cdots, 10\}.$$

(6) Sample space associated with the experiment of tossing a coin until a head appears for the first time and recording the number of tosses is :

$$S=\{1, 2, 3, \cdots\cdots, \infty\}.$$

Event.

A set of outcomes of an experiment is called *event*. Thus, an event is a subset of the sample space of an experiment.

Every set of sample points in a sample space is an event.

As the empty set ϕ and the universal set S are the subsets of S, ϕ and S are also events. The event ϕ is called *impossible* event and the event S is called *sure* or *certain* event.

Elementary Event :

An event A, in the sample space S, consisting *exactly* one sample point of S is called an *elementary* or *simple* event.

Thus an elementary event is an outcome of an experiment that cannot be decomposed into a combination of other elementary events of the sample space.

Compound Event :

An event is *compound* or *composite* if it can be decomposed into elementary events.

In an experiment of tossing two unbiased coins there are four elementary events, *i.e.*, HH, HT, TH and TT but the event 'one head and one tail' is a composite event consisting of two elementary events HT and TH.

Union.

The union of two events A and B is the set of all sample points *belonging to* A or B (or both) and is denoted by A ∪ B.

Symbolically, $A \cup B=\{x : x \in A \text{ or } x \in B\}$.

Intersection.

The intersection of two events A and B is the set of all sample points *common* to both A and B and is denoted by $A \cap B$.

Symbolically, $A \cap B = \{x : x \in A \text{ and } x \in B\}$.

Mutually Exclusive Events.

Events are *mutually exclusive* or *disjoint* when they have no points or elements in common.

Thus, if $A_1, A_2, \ldots\ldots, A_n$ be n mutually exclusive events, then

$$A_1 \cap A_2 \cap A_3 \cap \cdots\cdots \cap A_n = \phi.$$

So, for two disjoint events A and B, we have $A \cap B = \phi$.

Complement.

An event $\overline{A}$ is said to be the *complement* of an event A if $\overline{A}$ consists of all sample points in the sample space S that are not points of the event A.

Symbolically, $\overline{A} = \{x : x \in S, x \notin A\}$

The complement of an event A is also denoted by A' or A^c.

The number of all possible sample points or elementary events in the sample space of an experiment is denoted by $n(S)$ *and the number of sample points or elementary events in any event* A *is denoted by* $n(A)$.

Sample space of Equally Likely Outcomes : *Definition of Probability (classical definition)* :

Most often, the very physical nature of the experiment suggests that the different possible outcomes of the experiment are considered to be equally likely. In this case, various outcomes are assigned equal probabilities. Such a sample space, in which each and every sample point has the same probability, is called *equi-probable sample space* or *sample space of equally likely outcomes.*

(EXAMPLE : Experiment of throwing a die consists of 6 sample points in the sample space. Each of the sample points 'face 1', 'face 2',$\cdots\cdots$ 'face 6' has the probability $\frac{1}{6}$.)

Thus, if in a *finite* sample space of equally likely outcomes there are n sample points, the probability associated with each sample point would be $1/n$, as the sum of the probabilities of all the sample points in the given sample space must equal 1. Hence, for any event A, in the sample space S, consisting of m sample points, we have,

$$\begin{aligned} P(A) &= \frac{1}{n} + \frac{1}{n} + \cdots\cdots \text{ to } m \text{ terms} \\ &= \frac{m}{n} \\ &= \frac{\text{Number of sample points in A}}{\text{Number of sample points in S}} \\ &= \frac{\text{Number of elementary events in A}}{\text{Number of elementary events in S}} \\ &= \frac{n(A)}{n(S)}. \end{aligned}$$

Properties of P(A).

(i) $P(A) = \frac{n(A)}{n(S)}$ must always lie between 0 and 1, since $0 \leqslant n(A) \leqslant n(S)$.

(ii) $P(A) = 1$ when $n(A) = n(S)$, *i.e.*, when the event is *certain.*

(iii) $P(A) = 0$ when $n(A) = 0$, *i.e.*, when the event is *impossible.*

Note 1. To choose an object *at random* from n objects means that each object has the *same probability* $\frac{1}{n}$ of being chosen.

Note 2. To choose k objects *at random* from n objects ($k \leqslant n$) means that each set of k objects (disregarding order) has the *same probability* of being chosen as any other set of K objects.

Rules of Probability

I. *Rule of Addition :*

THEOREM 1 : *Theorem of Total Probability*—If $A_1, A_2, \ldots\ldots, A_m$ be m *mutually exclusive* events, then

$$P(A_1 \cup A_2 \cup \cdots\cdots \cup A_m) = P(A_1) + P(A_2) + \cdots\cdots + P(A_m)$$

i.e., the probability of A_1 or A_2 or ……or A_m is the sum of the probabilities of these events, provided the events are mutually exclusive.

PROOF :

Let a finite sample space S of equally likely outcomes of a random experiment consist of $n(S)$ sample points of which $n(A_1)$ sample points correspond to event A_1 and $n(A_2)$ sample points correspond to event A_2. Since the events A_1 and A_2 are mutually exclusive the number of sample points that correspond to either A_1 or A_2 is then $n(A_1)+n(A_2)$. Hence, by the definition of probability,

$$\begin{aligned} P(A_1 \cup A_2) &= \frac{n(A_1)+n(A_2)}{n(S)} \\ &= \frac{n(A_1)}{n(S)} + \frac{n(A_2)}{n(S)} \\ &= P(A_1)+P(A_2) \qquad \cdots \quad (1) \end{aligned}$$

Repeated application of (1) leads to

$$\begin{aligned} &P(A_1 \cup A_2 \cup A_3 \cup \cdots\cdots \cup A_{m-1} \cup A_m) \\ &\quad = P([A_1 \cup A_2 \cup A_3 \cup \cdots\cdots \cup A_{m-1}] \cup A_m) \\ &\quad = P(A_1 \cup A_2 \cup A_3 \cup \cdots\cdots \cup A_{m-1}) + P(A_m) \\ &\quad = P(A_1 \cup A_2 \cup A_3 \cup \cdots\cdots \cup A_{m-2}) + P(A_{m-1}) + P(A_m) \\ &\qquad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \\ &\quad = P(A_1)+P(A_2)+\cdots\cdots+P(A_m) \end{aligned}$$

when $A_1, A_2, \ldots\ldots, A_m$ are *mutually exclusive* events.

Cor. 1. If $A_1, A_2, \cdots\cdots, A_m$ are *mutually exclusive* events and $A = A_1 \cup A_2 \cup \cdots\cdots \cup A_m$, then $P(A) = P(A_1 \cup A_2 \cup \cdots\cdots \cup A_m)$

$$= P(A_1)+P(A_2)+\cdots+P(A_m).$$

Cor. 2. When the events $A_1, A_2, \cdots\cdots, A_m$ are *mutually exclusive* and also *exhaustive*, then $A_1 \cup A_2 \cup, \cdots\cdots, \cup A_m = S$ and we have $P(A_1 \cup A_2 \cup \cdots\cdots \cup A_m) = P(S) = 1$, implying that

$$P(A_1)+P(A_2)+\cdots+P(A_m)=1.$$

Cor. 3. The event A and its complement $\bar{A}$ are *mutually exclusive and hence*, $P(A \cup \bar{A}) = P(A) + P(\bar{A})$. Since $A \cup \bar{A} = S$, it follows that $P(A \cup \bar{A}) = P(S) = 1$. Therefore, $P(\bar{A}) = 1 - P(A)$.

Cor. 4. If A and B are two events then the events $(A \cup \bar{B})$ and $(A \cap B)$ are *mutually exclusive* and also, $A = (A \cap \bar{B}) \cup (A \cap B)$ and hence, $P(A) = P(A \cap \bar{B}) + P(A \cap B)$, implying $P(A \cap \bar{B}) = P(A) - P(A \cap B)$. Again, $(A \cap \bar{B}) = A - (A \cap B)$. So, $P\{A - (A \cap B)\} = P(A) - P(A \cap B)$.

Cor. 5. By De Morgan's Law, we have $(\bar{A} \cup B) = (A \cap B)$ and hence, $P(\bar{A} \cup \bar{B}) = P(A \overline{\cap B}) = 1 - P(A \cap B)$, by Cor. 3.

$\therefore \quad P(\bar{A} \cup \bar{B}) = 1 - P(A \cap B)$.

Cor. 6. If ϕ is an *impossible* event, then $P(\phi) = 0$. For, if A be any event, then we have, $A = A \cup \phi$. So, $P(A) = P(A \cup \phi) = P(A) + P(\phi)$, since A and ϕ are *mutually exclusive*, implying $P(\phi) = 0$.

THEOREM 2 :

Generalised Theorem on Total Probability—If A and B are any two events not necessarily mutually exclusive, in the sample space S then,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

PROOF :

Let $n(S)$ be the total number of sample points in a finite sample space S of a random experiment and $n(A \cup B)$ the number of sample points in $(A \cup B)$. Hence, by the definition of probability,

$$P(A \cup B) = \frac{n(A \cup B)}{n(S)}.$$

Now, $n(A \cup B) = n(A) + n(B) - n(A \cap B)$ [see page 402]

$$\therefore \quad P(A \cup B) = \frac{n(A) + n(B) - n(A \cap B)}{n(S)}$$
$$= \frac{n(A)}{n(S)} + \frac{n(B)}{n(S)} - \frac{n(A \cap B)}{n(S)}$$
$$= P(A) + P(B) - P(A \cap B).$$

The result can also be written as,

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B).$$

ALTERNATIVE PROOF :

Let A and B be any two events, in the sample space S of a random experiment, not necessarily mutually exclusive. Then $A - (A \cap B)$, $(A \cap B)$ and $B - (A \cap B)$ are three mutually exclusive events.

Now, $A = \{A - (A \cap B)\} \cup (A \cap B)$

$B = \{B - (A \cap B)\} \cup (A \cap B)$

and $(A \cup B) = \{A - (A \cap B)\} \cup (A \cap B) \cup \{B - (A \cap B)\}$

Hence, $P(A) = P\{A - (A \cap B)\} + P(A \cap B)$; ... (1)

since $A - (A \cap B)$ and $A \cap B$ are mutually exclusive.

$P(B) = P\{B - (A \cap B)\} + P(A \cap B)$; ... (2)

since, $B - (A \cap B)$ and $A \cap B$ are mutually exclusive.

and $P(A \cup B) = P\{A - (A \cap B)\} + P(A \cap B) + P\{B - (A \cap B)\}$ ... (3)

since, $A - (A \cap B)$, $A \cap B$ and $B - (A \cap B)$ are mutually exclusive.

From (3) with the help of (1) and (2), we get

$P(A \cup B) = P(A) - P(A \cap B) + P(A \cap B) + P(B) - P(A \cap B)$ [by Cor. 4.]

$= P(A) + P(B) - P(A \cap B)$

Theorem 3 :

If A, B and C are any three events, then

$$P(A\cup B\cup C)=P(A)+P(B)+P(C)-P(A\cap B)-P(A\cap C)-P(B\cap C)+P(A\cap B\cap C).$$

Proof :

If A, B and C are any three events, then

$$\begin{aligned}P(A\cup B\cup C)&=P[(A\cup B)\cup C]\\&=P(A\cup B)+P(C)-P[(A\cup B)\cap C]\\&=P(A)+P(B)-P(A\cap B)+P(C)-P[(A\cap C)\cup(B\cap C)]\end{aligned}$$

Since, by the distributive property of the sets,

$$(A\cup B)\cap C=(A\cap C)\cup(B\cap C).$$

Now, $P[(A\cap C)\cup(B\cap C)]=P(A\cap C)+P(B\cap C)-P[(A\cap C)\cap(B\cap C)]$

$$=P(A\cap C)+P(B\cap C)-P(A\cap B\cap C).$$

Since $(A\cap C)\cap(B\cap C)=(A\cap B\cap C)$.

Hence, $P(A\cup B\cup C)=P(A)+P(B)+P(C)-P(A\cap B)-P(A\cap C)-P(B\cap C)+P(A\cap B\cap C)$.

The above thorem can be extended to the case of more than three events.

For instance, if $A_1, A_2, \cdots\cdots, A_n$ be any n events, then it can be shown by induction that,

$$P(A_1\cup A_2\cup\cdots\cup A_n)=\sum_{i=1}^{n}P(A_i)-\sum_{\substack{i,j=1\\ i<j}}^{n}P(A_i\cup A_j)+\cdots\cdots$$
$$\cdots+(-1)^{n-1}P(A_1\cap A_2\cap\cdots A_n)\ \cdots\quad(1)$$

Note 1. When n events $A_1, A_2,\cdots\cdots, A_n$ are *mutually exclusive*, then from (1),

$$P(A_1\cup A_2\cup\cdots\cdots\cup A_n)=P(A_1)+P(A_2)+\cdots\cdots+P(A_n).$$

Since all other terms of R.H.S. (1) becomes zero.

Note 2. $P(A\cup B)=P(A)+P(B)-P(A\cap B)$

$\Rightarrow P(A\cup B)\leq P(A)+P(B)$, since $P(A\cap B)\geqslant 0$.

The sign of equality holds only when A and B are mutually exclusive. This inequality is known as *Boole's inequality*.

EXAMPLE :

Find the probability of throwing a total of 8 in tossing two balanced dice.

SOLUTION :

The sample space here consists of 36 sample points.

Let A denote the event of throwing a total of 8, then A consists of 5 sample points (2, 6), (3, 5), (4, 4), (5, 3), (6, 2).

$\therefore \quad P(A)=\frac{5}{36}$.

EXAMPLE :

A card is drawn from a well-shuffled pack of 52 cards. What is the probability of the card being either black or an ace ?

SOLUTION :

The sample space with the experiment of drawing a card from a well-shuffled pack of 52 cards consists of 52 sample points.

Let A denotes the event that the card is black, then A consists of 26 sample points since there are 26 black cards. If B denotes the event that the card is an ace, then B consists of 4 sample points, since the number of ace cards is 4. Now, $(A \cap B)$ consists of two sample points, viz. two black aces.

$\therefore \quad P(A)=\frac{26}{52}$, $P(B)=\frac{4}{52}$ and $P(A \cap B)=\frac{2}{52}$

$\therefore$ the probability that the card drawn is either black or an ace is $P(A \text{ or } B)=P(A \cup B)$. Since the events A and B are not mutually exclusive, we have,

$$\begin{aligned} P(A \cup B) &= P(A)+P(B)-P(A \cap B) \\ &= \tfrac{26}{52}+\tfrac{4}{52}-\tfrac{2}{52} \\ &= \tfrac{28}{52}=\tfrac{7}{13}. \end{aligned}$$

EXAMPLE :

An urn contains 13 balls numbering from 1 to 13. Find the probability that a ball selected at random is a ball with number that is a multiple of 3 or 4.

SOLUTION :

Here, the sample space consists of 13 sample points. If A be the event that the ball selected is a ball with number that is a multiple of 3, then A consists of 4 sample points, viz. 3, 6, 9, 12.

$$\therefore \quad P(A)=\tfrac{4}{13}.$$

Similarly, if B be the event that the ball selected is a ball with number that is a multiple of 4, then B consists of 3 sample points, viz. 4, 8 12.

$$\therefore \quad P(B)=\tfrac{3}{13}.$$

The event $A \cap B$ comprises of only one sample point, *i.e.*, 12 which is a multiple of both 3 and 4.

$$\therefore \quad P(A \cap B)=\tfrac{1}{13}\,;$$

$$P(A \cup B)=P(A)+P(B)-P(A \cap B)$$

$=\tfrac{4}{13}+\tfrac{3}{13}-\tfrac{1}{13}=\tfrac{6}{13}$ which is the probability that the ball selected at random is a multiple of 3 or 4.

EXAMPLE :

In drawing a card from a well-shuffled pack, find the probability that the card drawn will be either a spade or a heart.

SOLUTION :

The experiment of drawing a card from a well-shuffled pack consists of 52 equally likely outcomes. So the sample space consists of 52 sample points.

Let $A=\{$the card is a spade$\}$,

and $B=\{$the card is a heart$\}$.

$\therefore$ A consists of 13 sample points, since there are 13 spade cards.

$$\therefore \quad P(A)=\tfrac{13}{52}.$$

Similarly, B also consists of 13 sample points.

$$\therefore \quad P(B)=\tfrac{13}{52}.$$

The events A and B are mutually exclusive, since a spade and a heart cannot both occure in the same draw.

$$\therefore \quad P(A \cup B)=P(A)+P(B)=\tfrac{13}{52}+\tfrac{13}{52}=\tfrac{1}{2}.$$

EXAMPLE :

The probability that a contractor will get a plumbing contract is $\tfrac{2}{3}$, and the probability that he will not get an electric contract is $\tfrac{5}{9}$. If the probability of getting at least one contract is $\tfrac{4}{5}$, what is the probability that he will get both the contracts ? (C. A. 1979)

SOLUTION :

Let A be the event that the contractor will get a plumbing contract, then $P(A)=\tfrac{2}{3}$.

If B be the event that the contractor will get an electric contract, then $P(\bar{B}) = \frac{5}{9}$ (given).

$\therefore \quad P(B) = 1 - P(\bar{B}) = 1 - \frac{5}{9} = \frac{4}{9}$.

Also given, $P(A \cup B) = \frac{4}{5}$.

$\therefore$ by the addition rule, we have,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

$\therefore \quad \frac{4}{5} = \frac{2}{3} + \frac{4}{9} - P(A \cap B)$,

or, $P(A \cap B) = \frac{2}{3} + \frac{4}{9} - \frac{4}{5} = \frac{14}{45}$.

$\therefore$ Probability that the contractor will get both the contracts $= \frac{14}{45}$.

Example :

A and B are two events, not mutually exclusive, connected with a random experiment E. If $P(A) = \frac{1}{4}$, $P(B) = \frac{2}{5}$ and $P(A \cup B) = \frac{1}{2}$, find the values of the following probabilities :

(i) $P(A \cap B)$, (ii) $P(A \cap B^c)$, (iii) $P(A^c \cup B^c)$

where c stands for the complement. (C. U. 1980)

Solution :

(i) We have,

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$, since A and B are not mutually exclusive.

i.e., $\frac{1}{2} = \frac{1}{4} + \frac{2}{5} - P(A \cap B)$

or $P(A \cap B) = \frac{1}{4} + \frac{2}{5} - \frac{1}{2} = \frac{3}{20}$.

(ii) $P(A \cap B^c) = P(A) - P(A \cap B) = \frac{1}{4} - \frac{3}{20} = \frac{2}{20} = \frac{1}{10}$ [by Cor. (4)]

(iii) $P(A^c \cup B^c) = P\{(A \cap B)^c\} = 1 - P(A \cap B) = 1 - \frac{3}{20} = \frac{17}{20}$ [by Cor. (5)]

Compound Probability.

The probability of occurrence of two or more events simultaneously is termed as compound probability. The usual notation for compound probability for two events A_1, A_2 is $P(A_1 \cap A_2)$ and for n events $A_1, A_2, \cdots\cdots, A_n$ is $P(A_1 \cap A_2 \cap \cdots \cap A_n)$.

Note. Compound probability is also called *Joint Probability*.

Conditional Probability.

Let A and B be two events in the sample space S of a random experiment, such that $P(A) > 0$. The probability of occurrence of event B, subject to the condition that event A has already occurred, is called the conditional probability of B, given A. In terms of symbols, it is written $P(B|A)$ and is defined by the proportion of sample points in event B among the sample points in event A.

The vertical bar is read 'given'.

II. *Rule of Multiplication :*

THEOREM 1 :

The probability of simultaneons occurrence of two events is equal to the product of the probability of one of the events by the conditional probability of the other, given that the first one has already occurred.

PROOF :

Let S be a finite sample space of equally likely outcomes of a random experiment and $n(S)$ be the total number of sample points in S. Let A be any event in S with sample points $n(A)$, such that $n(A) > 0$. Together with the sample points $n(A)$, let us consider, the number of sample points $n(A \cap B)$ that are simultaneously in A and B. Then the ratio

$$\frac{n(A \cap B)}{n(A)}$$

is the proportion of sample points in B among the sample points in A and is the conditional probability of B, given A.

$$\therefore \quad P(B|A) = \frac{n(A \cap B)}{n(A)}.$$

$$= \frac{n(A \cap B)}{n(S)} \cdot \frac{n(S)}{n(A)}$$

$$= \frac{n(A \cap B)}{n(S)} \Big/ \frac{n(A)}{n(S)}$$

$$= P(A \cap B)/P(A). \qquad \cdots (1)$$

In the similar way we can derive,

$$P(A|B) = P(A \cap B)/P(B), \text{ if } n(B) > 0, \textit{ i.e.}, P(B) > 0 \quad \cdots \quad (2)$$

From the equations (1) and (2), we get,

$$P(A \cap B) = P(A)\,P(B|A) = P(B)\,P(A|B) \quad \cdots \quad (3)$$

Note 1. In computing P (B|A) we are essentially computing the probability of B with reference to the subset A of the original sample space S, rather than with reference to the original sample space S. This subset A of S is called the *reduced sample space*. Thus in the conditional probability of B, given A, the reduced sample space is A, the given event. But in computing P (B) we compute the probability of B with reference to the original sample space S. P (B) is really an abbreviation for P (B|S). But S is dropped as *understood*.

Note 2. P (B/A) may be computed either directly by calculating the probability of B with reference to the reduced sample space A or by using

$$P(B|A)=P(A\cap B)|P(A)$$

where P (A∩B) and P (A) are computed with reference to the original sample space S.

Remark. In finite sample space of equally likely outcomes, P (A) should be not equal to zero to define P (B/A), and P (B) should not be equal to zero to define P (A/B).

THEOREM 2 :

Extention of Theorem of Compound Probability—The probability of simultaneons occurrence of n events A_1, A_2, ······, A_n is

$$P(A_1\cap A_2\cap\cdots\cap A_n)=P(A_1).\ P(A_2|A_1).\ P(A_3|A_1\cap A_2)\cdots\cdots$$
$$\cdots\cdots P(A_n|A_1\cap A_2\cap\cdots\cap A_{n-1})$$

provided $P(A_1)>0$, $P(A_2|A_1)>0$, $P(A_3|A_1\cap A_2)>0$, ······

$$\cdots\cdots P(A_{n-1}/A_1\cap A_2\cap\cdots\cap A_{n-2})>0.$$

PROOF :

Let n (S) be the total number of sample points in a finite sample space S of a random experiment of which there are $n(A_1)$ sample points in A_1 and $n(A_1\cap A_2)$ sample points in $(A_1\cap A_2)$.

$$\therefore\quad P(A_1\cap A_2)=\frac{n(A_1\cap A_2)}{n(S)}\qquad\cdots\quad(1)$$

Since $P(A_1)>0$, *i.e.*, $n(A_1)>0$, equation (1) can be written in the form :

$$P(A_1\cap A_2)=\frac{n(A_1\cap A_2)}{n(A_1)}.\ \frac{n(A_1)}{n(S)}.$$

But $\quad \dfrac{n(A_1\cap A_2)}{n(A_1)}=P(A_2|A_1)$ and $\dfrac{n(A_1)}{n(S)}=P(A_1)$.

Hence, $\quad P(A_1\cap A_2)=P(A_1)\ P(A_2|A_1)$.

Again, since $P(A_1) > 0$ and $P(A_2|A_1) > 0$, we have from (2), $P(A_1 \cap A_2) > 0$, and hence,

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P\{(A_1 \cap A_2) \cap A_3\} \\ &= P(A_1 \cap A_2)\, P(A_3/A_1 \cap A_2) \\ &= P(A).P(A_2 | A_1).P(A_3 | A_1 \cap A_2) \end{aligned}$$

Proceeding in this way we get,

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1).P(A_2 | A_1) \cdots P(A_n | A_1 \cap A_2 \cap \cdots \cap A_{n-1})$$

Cor. 1. If the events $A_1, A_2, \ldots\ldots, A_n$ are mutually exclusive and exhaustive, then $B \cap A_1$, $B \cap A_2$, ..., $B \cap A_n$ are mutually exclusive and $B = (B \cap A_1) \cup (B \cap A_2) \cup \cdots\cdots \cup (B \cap A_n)$.

$$\begin{aligned} \therefore \quad P(B) &= P\{(B \cap A_1) \cup (B \cap A_2) \cup \cdots\cdots \cup (B \cap A_n)\} \\ &= P(B \cap A_1) + P(B \cap A_2) + \cdots\cdots + P(B \cap A_n) \\ &= P(A_1)P(B | A_1) + P(A_2)P(B | A_2) + \cdots\cdots + P(A_n)P(B/A_n), \end{aligned}$$

provided $P(A_1) > 0$, $P(A_2) > 0$, ..., $P(A_n) > 0$.

Cor. 2. If A and B be two events, then $A = (A \cap B) \cup (A \cap \bar{B})$ and $A \cap B$ and $A \cap \bar{B}$ are mutually exclusive.

Hence $P(A) = P(A \cap B) + P(A \cap \bar{B})$.

Independent Events.

Two events are said to be *independent* if the occurrence of one event does not influence the occurrence of the other event.

EXAMPLE :

Successive tosses of a fair coin are independent. If a fair coin is tossed twice, the event 'Head' in the first toss and the event 'Head' in the second toss are independent since the occurrence of 'Head' in any toss does not influence the occurrence of 'Head' of the other toss and the probability of getting a Head, say, in the second toss, which is $\frac{1}{2}$, does not change, if it be known that the first toss has resulted in a 'Head' or not.

Similarly, if two cards are drawn *with replacement* from a pack of well-shuffled cards, the events (A) 'black card in the first draw' and (B) 'black card in the second draw' are independent. But if the drawing is made *without replacement*, the events A and B will be dependent events.

Two or more events are considered to be independent if the occurrence or non-occurrence of one has no effect on whether the other or others occur.

If events A and B are such that

$$P(A/B) = P(A)$$

then the occurrence of event B does not alter the probability of event A and hence we say that the event A is *independent* of event B.

If event A is independent of event B, then the event B is also independent of event A.

For, we have

$$P(A \cap B) = P(A)P(B \mid A) = P(B)\,P(A \mid B)$$

Now, since event A is independent of event B,

$$P(A \mid B) = P(A).$$

$$\therefore \quad P(A)P(B \mid A) = P(B)P(A)$$

$$\Rightarrow \qquad P(B \mid A) = P(B), \quad \text{since } P(A) > 0$$

that is, the event B is independent of event A.

For independent events A and B the compound probability theorem takes the following simple form

$$P(A \cap B) = P(A)P(B).$$

This relation is also used to define the independence of the two events A and B.

Definition : Two events A and B are called *independent* when the relation $P(A \cap B) = P(A)P(B)$ holds otherwise they are called *dependent* events.

Remark : The relation $P(A \cap B) = P(A)P(B)$ is accepted to remain valid for all values of P(A) and P(B) including $P(A) = 0$ and $P(B) = 0$.

In general, when a finite number of events $A_1, A_2, \ldots\ldots, A_n$ are independent, we have

$$P(A_1 \cap A_2 \cap \cdots\cdots \cap A_n) = P(A_1)P(A_2)\cdots\cdots P(A_n)$$

Note. If two events A and B are mutually exclusive then $A \cap B = \phi$ and hence $P(A \cap B) = P(\phi) = 0$.

THEOREM :

If A and B are two independent events, then $\bar{A}$ and $\bar{B}$ are also independent.

PROOF : We have

$$\begin{aligned} P(\bar{A} \cap \bar{B}) &= P(\overline{A \cup B}), \text{ by De Morgan's Law} \\ &= 1 - P(A \cup B) \\ &= 1 - \{P(A) + P(B) - P(A \cap B)\} \\ &= 1 - P(A) - P(B) + P(A \cap B) \\ &= 1 - P(A) - P(B) + P(A)\,P(B), \end{aligned}$$

since A and B are independent.

$$=1-P(A)-P(B)\{1-P(A)\}$$
$$=\{1-P(A)\}\{1-P(B)\}$$
$$=P(\bar{A})\,P(\bar{B}),$$

that is, the events $\bar{A}$ and B are independent.

Theorem :

If A and B are two independent events, then A and $\bar{B}$ are also independent.

Proof :

Since $(A\cap\bar{B})=A-(A\cap B)$, we have,

$$P(A\cap\bar{B})=P\{A-(A\cap B)\}$$
$$=P(A)-P(A\cap B) \quad \text{[by Cor. 4—see page 425]}$$
$$=P(A)-P(A).P(B), \quad \text{since A and B are independent events}$$
$$=P(A)\,\{1-P(B)\}$$
$$=P(A)\,.\,P(\bar{B}),$$

that is, the events A and $\bar{B}$ are independents.

Alternative Proof :

We have, indeed,

$$P(B\,|\,A)+P(\bar{B}\,|\,A)=1$$
$$\Rightarrow P(B)+P(\bar{B}\,|\,A)=1, \text{ since A and B are independent}$$
$$\Rightarrow P(\bar{B}\,|\,A)=1-P(B)=P(\bar{B}),$$

that is, the events A and $\bar{B}$ are independent.

Theorem :

If A and B are two independent events, then events $\bar{A}$ and B are also independent.

Proof :

Since $(\bar{A}\cap B)=B-(A\cap B)$, we have,

$$P(\bar{A}\cap B)=P\{B-(A\cap B)$$
$$=P(B)-P(A\cap B) \quad \text{[by Cor. 4—see page 425]}$$
$$=P(B)-P(A).P(B), \quad \text{since A and B are independent events}$$
$$=P(B)\{1-P(A)\}$$
$$=P(B)\,P(\bar{A}),$$

that is, the events $\bar{A}$ and B are independent.

ALTERNATIVE PROOF :

We have, indeed,

$$P(A|B)+P(\bar{A}|B)=1$$
$$\Rightarrow P(A)+P(\bar{A}|B)=1, \text{ since A and B are independent}$$
$$\Rightarrow P(\bar{A}|B)=1-P(A)=P(\bar{A}),$$

that is, the events $\bar{A}$, B are independent.

THEOREM :

If A and B are two independent events, then

$$P(A \cup B)=1-P(\bar{A}).P(\bar{B}).$$

PROOF :

Since $(\overline{A \cup B})$ is the complement of $(A \cup B)$, we have

$$\begin{aligned} P(A \cup B) &= 1-P(\overline{A \cup B}) \\ &= 1-P(\bar{A} \cap \bar{B}) \quad [\text{by De Morgan's Law}] \\ &= 1-P(\bar{A}).P(\bar{B}) \end{aligned}$$

[Since A and B are independent, $\bar{A}$ and $\bar{B}$ are also independent.]

Note. If there are n independent events $A_1, A_2, \cdots A_n$ then

$$P(A_1 \cap A_2 \cap \cdots \cap A_n)=1-P(\bar{A}_1)\, P(\bar{A}_2) \cdots P(\bar{A}_n).$$

THEOREM :

If A and B are two independent events such that $P(A) > 0$ and $P(B) > 0$, then $(A \cap B) \neq \phi$, *i.e.*, A and B are not mutually exclusive.

PROOF :

If possible, let $(A \cap B)=\phi$, where ϕ is an impossible event. Then

$$P(A \cap B)=P(\phi)=0$$

Since, A and B are independent events,

$$P(A \cap B)=P(A).P(B)$$

Therefore, $P(A).P(B)=0$ implying that either $P(A)=0$ or $P(B)=0$ which contradicts the hypothesis of the theorem that $P(A) \neq 0$ and $P(B) \neq 0$. Hence $(A \cap B) \neq \phi$, *i.e.*, A and B are not mutually exclusive.

Thus, two events with non-zero probabilities cannot be mutually exclusive and independent simultaneously.

Two events A and B can be mutually exclusive and independent simultaneously if either $P(A)=0$ or $P(B)=0$.

THEOREM :

If two events A and B, having non-zero probabilities, are mutually exclusive then both P(A|B) and P(B|A) are equal to zero.

PROOF :

Since events A and B are mutually exclusive, we have,

$$P(A \cap B) = P(\phi) = 0.$$

Now, $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0}{P(B)} = 0$, since $P(B) > 0$

Similarly, $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0}{P(A)} = 0$, since $P(A) > 0$.

Baye's Thorem.

Let A_1, A_2, ... A_n be n mutually exclusive events whose union is the sample space S in a random experiment and let B be an arbitrary event in the sample space such that $P(B) \neq 0$, then,

$$P(A_i|B) = \frac{P(A_i).P(B|A_i)}{\sum_{j=1}^{n} P(A_j).P(B|A_j)}$$

PROOF :

By the law of Conditional Probability,

$$P(B|A_i) = \frac{P(A_i \cap B)}{P(A_i)}$$

or $\quad P(A_i \cap B) = P(A_i).P(B|A_i)$

Similarly, $P(B \cap A_i) = P(B).P(A_i|B)$

Now, since $P(A_i \cap B) = P(B \cap A_i)$, it follows that

$$P(A_i)\ P(B|A_i) = P(B).P(A_i|B)$$

$$\therefore \quad P(A_i|B) = \frac{P(A_i).P(B|A_i)}{P(B)}$$

Since the events A_1, A_2, ..., A_n are mutually exclusive and exhaustive, and $P(B) \neq 0$,

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + \cdots + P(A_n \cap B)$$
$$= P(A_1)\ P(B|A_1) + P(A_2)\ P(B|A_2) + \cdots + P(A_n).P(B|A_n)$$

$$\therefore \quad P(A_i/B) = \frac{P(A_i).P(B|A_i)}{P(A_1).P(B|A_1) + P(A_2).P(B|A_2) + \cdots + P(A_n).P(B|A_n)}$$

EXAMPLE :

An urn contains 7 black and 5 white balls. Two balls are drawn at random one after the other. Find the probability that both balls drawn are black if (i) when first ball drawn is not replaced before drawing the second and (ii) when first ball drawn is replaced before drawing the second ball.

SOLUTION :

(i) The sample space consists of 12 sample points as there are altogether $7+5=12$ balls.

Let $A=\{\text{1st ball drawn is black}\}$,

and $B=\{\text{2nd ball drawn is black}\}$.

The event A consists of 7 sample points as there are 7 black balls.

$$\therefore \quad P(A)=\tfrac{7}{12}.$$

Now since the first ball drawn is black and is not replaced the sample space reduces to 11 points only as there are only 6 black and 5 white balls left. The event B now consists of 6 sample points as there are now 6 black balls.

$$\therefore \quad P(B|A)=\tfrac{6}{11}.$$

$\therefore$ the probability that both balls drawn black is,

$$P(A\cap B)=P(A).P(B|A)$$
$$=\tfrac{7}{12}\times\tfrac{6}{11}=\tfrac{7}{22}.$$

(ii) Now the events A and B are independent,

$$P(A)=\tfrac{7}{12}\ ;\ P(B)=\tfrac{7}{12}\ ;$$

and $P(A\cap B)=P(A).P(B)=\tfrac{7}{12}\times\tfrac{7}{12}=\tfrac{49}{144}.$

EXAMPLE :

A bag contains 7 red and 5 white balls. 2 balls are drawn at random without replacement. What is the probability that the second ball is red, knowing that the first ball is red ?

SOLUTION :

Let $A=\{\text{1st ball is red}\}$; $B=\{\text{2nd ball is red}\}$.

Total number of balls in the bag $=7+5=12$.

$\therefore$ There are ${}^{12}C_2$ ways of drawing 2 balls from the bag and hence the sample space consists of ${}^{12}C_2$ sample points.

The number of ways of drawing 2 red balls from 7 red balls is ${}^{7}C_2$. So,

$$P(A\cap B)=\frac{{}^{7}C_2}{{}^{12}C_2}=\frac{7.6}{12.11}=\frac{7}{22}.$$

P(A) = probability that the 1st ball drawn is red $= \frac{7}{12}$.

Now, $P(A \cap B) = P(A).P(B|A)$

$\therefore \quad \frac{7}{22} = \frac{7}{12}.P(B|A)$

$\Rightarrow P(B|A) = \frac{7}{22} \times \frac{12}{7} = \frac{6}{11}$.

Note. P(B|A) can be directly computed. For when the 1st ball drawn is red, there remain 6 red and 5 white balls in the bag and hence $P(B|A) = \frac{6}{6+5} = \frac{6}{11}$.

EXAMPLE :

In throwing two fair-dice, find the probability of getting a total of 9, when it is known that second dice will show a smaller value than the first dice.

SOLUTION :

The sample space consists of $6^2 = 36$ sample points.

Let A = {Sum of the points in two die = 9},

B = {Points in the 1st dice is greater than the points in the 2nd dice in any throw}.

Thus A = {(6, 3), (5, 4), (4, 5), (3, 6)} ;

B = {(2, 1), (3, 1), (4, 1), (5, 1), (6, 1), (3, 2), (4, 2), (5, 2), (6, 2), (4, 3), (5, 3), (6, 3), (5, 4), (6, 4), (6, 5)}.

So, event A consists of four and event B consists of fifteen sample points.

Again $(B \cap A)$ consists of only two sample points. viz. (6, 3) and (5, 4).

$\therefore \quad P(B \cap A) = \frac{2}{36} = \frac{1}{18}$

Also, $P(B) = \frac{15}{36}$

Now, $P(B \cap A) = P(B).P(A|B)$.

$\therefore \quad \frac{1}{18} = \frac{15}{36}.P(A|B)$

$\Rightarrow A(A|B) = \frac{1}{18} \times \frac{36}{15} = \frac{2}{15}$.

EXAMPLE :

Two persons X and Y appear in an interview for two vacancies in the same post. The probability of X's selection is $\frac{1}{5}$ and that of Y's selection is $\frac{1}{3}$, what is the probability that

(i) both X and Y will be selected,

(ii) only one of them will be selected,

and (iii) none of them will be selected.

Solution :

Let $A=\{X \text{ will be selected}\}$,

and $B=\{Y \text{ will be selected}\}$.

(i) In this case the events A and B are independent.

$\therefore \quad P(A\cap B)=P(A).P(B)=\frac{1}{5}\times\frac{1}{3}=\frac{1}{15}$

$\therefore$ Probability that both X and Y will be selected $=\frac{1}{15}$.

(ii) The event A and $\overline{B}$ are independent and so also the events $\overline{A}$ and B, since events A and B are independent.

$\therefore$ P(only one of them will be selected)

$$\begin{aligned}
&=P(A\cap\overline{B})+P(\overline{A}\cap B)\\
&=P(A).P(\overline{B})+P(\overline{A}).P(B)\\
&=P(A).\{1-P(B)\}+\{1-P(A)\}.P(B)\\
&=\tfrac{1}{5}.(1-\tfrac{1}{3})+(1-\tfrac{1}{5}).\tfrac{1}{3}\\
&=\tfrac{1}{5}.\tfrac{2}{3}+\tfrac{4}{5}.\tfrac{1}{3}\\
&=\tfrac{6}{15}=\tfrac{2}{5}.
\end{aligned}$$

(iii) The events $\overline{A}$ and $\overline{B}$ are independent, since the events A and B are independent.

$\therefore$ P (none of X and Y will be selected)

$$\begin{aligned}
&=P(\overline{A}\cap\overline{B})\\
&=P(\overline{A}).P(\overline{B})\\
&=[1-P(A)].[1-P(B)]\\
&=(1-\tfrac{1}{5}).(1-\tfrac{1}{3})\\
&=\tfrac{4}{5}.\tfrac{2}{3}=\tfrac{8}{15}.
\end{aligned}$$

Example :

Three bags contain repectively 5 white, 3 black balls ; 7 white, 8 black balls ; 4 white, 5 black balls. One bag is chosen at random and a ball from it is also chosen at random. What is the probability that the ball is white ?

Solution :

Let $A=\{\text{ball drawn is white}\}$,

and $B_i=\{i\text{-th bag is chosen}\}$; $i=1, 2, 3$.

The events B_1, B_2, B_3 are mutually exclusive exhaustive and also none of them is impossible.

Then $P(B_1)=\frac{1}{3}$, $P(B_2)=\frac{1}{3}$ and $P(B_3)=\frac{1}{3}$

and $P(A|B_1)=\dfrac{5}{5+8}$, $P(A|B_2)=\frac{7}{15}$ and $P(A|B_3)=\frac{4}{9}$.

$$\begin{aligned}\therefore\quad P(A) &= P(B_1).\ P(A|B_1)+P(B_2).\ P(A|B_2)+P(B_3).\ P(A|B_3)\\ &= \tfrac{1}{3}.\ \tfrac{5}{13}+\tfrac{1}{3}.\ \tfrac{7}{15}+\tfrac{1}{3}.\ \tfrac{4}{9}\\ &= \tfrac{1}{3}\left(\tfrac{5}{13}+\tfrac{7}{15}+\tfrac{4}{9}\right)\\ &= \tfrac{758}{1755}.\end{aligned}$$

EXAMPLE :

There are two men aged 30 and 36 years. The probability to live 35 years more is ˙67 for the 30 years old and ˙60 for the 36 years old person. Find the probability that at least one of these persons will be alive 35 years hence.

SOLUTION :

Let A be the event that 30 years old person will die within 35 years and B be the event that 36 years old person will die within 35 yearss.

$$\begin{aligned}\therefore\quad P(A) &= 1-{}^{\cdot}67={}^{\cdot}33\\ P(B) &= 1-{}^{\cdot}60={}^{\cdot}40.\end{aligned}$$

Since the events A and B are independent, the probability that both persons will die within 35 years is given by :

$$P(A\cap B)=P(A).P(B)={}^{\cdot}33\times{}^{\cdot}40={}^{\cdot}132$$

$\therefore$ the probability that at least one of the persons will be alive 35 years hence is,

$$1-P(A\cap B)=1-0{}^{\cdot}132=0{}^{\cdot}968.$$

EXAMPLE :

Urn-1 contain 5 red and 5 black balls, urn-2 contains 4 red and 8 black balls and urn-3 contains 3 red and 6 black balls. One urn is chosen at random and a ball is drawn. The colour of the ball is black. What is the probability that it has been drawn from urn-3. ?

SOLUTION :

Let $A=\{\text{A black ball is drawn}\}$

and $B_i=\{i\text{-th urn is chosen}\}$; $i=1, 2, 3$

Then, $P(B_1)=\frac{1}{3}$, $P(B_2)=\frac{1}{3}$ and $P(B_3)=\frac{1}{3}$;

$P(A|B_1)=\frac{5}{10}$; $P(A|B_2)=\frac{8}{12}$ and $P(A|B_3)=\frac{6}{9}$.

$\therefore$ P(urn-3 is chosen/ball drawn is black)

$$=P(B_3/A)$$

$$=\frac{P(B_3).\ P(A\,|\,B_3)}{P(B_1).\ P(A\,|\,B_1)+P(B_2).\ P(A\,|\,B_2)+P(B_3).\ P(A\,|\,B_3)}$$

$$=\frac{\frac{1}{3}.\ \frac{6}{9}}{\frac{1}{3}.\ \frac{5}{10}+\frac{1}{3}.\ \frac{6}{12}+\frac{1}{3}.\ \frac{6}{9}}$$

$$=\frac{\frac{2}{9}}{\frac{1}{6}+\frac{2}{9}+\frac{2}{9}}=\frac{2}{9}\times\frac{18}{11}=\frac{4}{11}.$$

Example :

A fair coin is tossed twice, show that the events,

A = {Head on the first coin},

B = {Head on the second coin},

C = {Head on one coin only},

are pairwise independent but not independent themselves.

Solution :

Here, the sample space contains four sample points, viz. HH, HT, TH and TT, the event A contains 2 sample points, HH and HT ; B contains two points, HH and TH and C also contains two points, HT, TH.

Therefore, we have,

$$P(A)=\tfrac{2}{4}=\tfrac{1}{2},\ P(B)=\tfrac{2}{4}=\tfrac{1}{2},\ P(C)=\tfrac{2}{4}=\tfrac{1}{2}.$$

$$\therefore\ P(A).P(B)=\tfrac{1}{4},\ P(B).P(C)=\tfrac{1}{4} \text{ and } P(C).P(A)=\tfrac{1}{4}$$

Also, $A\cap B$ contains one sample point, HH, $A\cap C$ contains one point, HT, and $B\cap C$ contains one point TH.

$$\therefore\ P(A\cap B)=\tfrac{1}{4},\ P(A\cap C)=\tfrac{1}{4} \text{ and } P(B\cap C)=\tfrac{1}{4}$$

implying $P(A\cap B)=P(A).P(B),\ P(A\cap C)=P(A).P(C)$

and $P(B\cap C)=P(B).P(C).$

that is, events A, B and C are pairwise independent.

But $P(A\cap B\cap C)=(P\phi)=0$ and $P(A).P(B).P(C)=\frac{1}{8}$

$$\therefore\ P(A\cap B\cap C)\neq P(A).P(B).P(C),$$

that is, the events A, B, C are not independent.

Repeated Trials.

Let a random experiment has two outcomes—the occurrence of an event A, called a *success* or its non-occurrence a *failure*. If the probability of success in any trial is p and the probability of failure in any trial is $q=1-p$, then the probability of r successes in n *independent* trials is given by

$$P_n(r)={}^nC_r\, p^r q^{n-r}.$$

For, the probability that the first r trials produce successes at each trial and the remaining $(n-r)$ trials produce only failures at each trial by the theorem of Compound Probability is

$$p.\, p \cdots \cdots p.\, q.\, q \cdots \cdots q=p^r q^{n-r}$$

since there are r factor p and $(n-r)$ factors q.

Similarly, the probability for any fixed sequence of r successes and $(n-r)$ failures is $p^r q^{n-r}$. The number of such sequences is ${}^nC_r$, for we must choose exactly r positions out of n for the successes and the remaining $(n-r)$ positions for the failures. Since these ${}^nC_r$ outcomes are *mutually exclusive*, the probability of r successes in n independent trials, by the theorem of addition of probabilities, is ${}^nC_r p^r q^{n-r}$.

Note. The observations of a random experiment are often referred to as *trials.*

EXAMPLE :

Find the probability of getting 4 heads and 3 tails in tossing 7 coins.

SOLUTION :

Assuming that the coins are fair, tossing of 7 coins is the same as tossing a coin 7 times.

Let the event 'appearance of a head' be called a *success.*

Now, Probability of getting a head $=\frac{1}{2}$.

$\therefore$ the probability of success in each trial $=p=\frac{1}{2}$.

$\therefore$ $q=1-p=1-\frac{1}{2}=\frac{1}{2}$.

Here the trials are independent, since p or q is not affected by the results of any other tossing.

Therefore the probability that there will be 4 heads (and hence 3 tails) in 7 trials is,

$${}^7C_4\,(\tfrac{1}{2})^4.(\tfrac{1}{2})^3=\frac{7.6.5.4}{4.3.2.1}.\frac{1}{2^7}=\frac{35}{128}.$$

EXAMPLE :

A factory produces articles among which 20% are defective. If 5 articles are selected at random from a day's production, find the probability that there will be exactly 2 defectives.

SOLUTION :

Let the event 'occurrence' of a defective article be called a success. Then,

$$p = \text{probability that an article is defective} = \tfrac{20}{100} = \cdot 20$$

$$\therefore \quad q = 1 - \cdot 20 = \cdot 80.$$

Here, the trials are independent, since the probability of occurrence of defective article in any trial is not affected by the occurrence or non-occurrence of a defective article in any other trial.

Therefore, the probability that there will be exactly 2 defectives

$$= {}^5C_2(\cdot 2)^2(\cdot 8)^3 = 0\cdot 2048.$$

EXAMPLE :

Find the probability of having 3 boys in a family with 5 children.

SOLUTION :

Let the event of having a boy be called a success. Then,

$$p = \text{probability of having a boy} = \tfrac{1}{2}.$$

$$\therefore \quad q = 1 - \tfrac{1}{2} = \tfrac{1}{2}.$$

Here, the trials are independent since the probability of having a boy or a girl is not affected by the result of any other trial.

Therefore, the probability that there will be 3 boys (and hence 2 girls) in a family of 5 children is

$${}^5C_3(\tfrac{1}{2})^3(\tfrac{1}{2})^{5-3}$$

$$= \tfrac{5 \times 4 \times 3}{3 \times 2 \times 1} \cdot \tfrac{1}{8} \cdot \tfrac{1}{4}$$

$$= \tfrac{5}{16}.$$

Random Variable

A variable whose value is a numerical quantity determined by the outcome of a random experiment is called a *random* variable.

EXAMPLE :

The experiment of tossing a number of coins, say 3, has eight possible outcomes HHH, HHT, THH, HTH, HTT, THT, TTH, TTT but these outcomes are not numerical. We may, however, associate the numbers 0, 1, 2, 3 corresponding to the four possibilities regarding the number of heads that appear in 3 coins. If we now let the variable x represent the number of heads observed in the experiment, then the possible values that x can have is 0, 1, 2, 3. Since the value of the variable x is a number determined by the outcome of an experiment, it is a random variable.

Random variables are of two types : (i) continuous and (ii) discrete.

Discrete Random Variable :

When a random variable can assume only the values that can be counted, is called a *discrete* random variable.

Examples of discrete random variables are the number of defective items in a lot, number of accidents in a month, etc.

Continuous Random Variable :

A random variable is *continuous* if it can assume any value within a given range. It has infinite possible values.

Examples of continuous random variables are average cost per unit of an item, weight of students in a class, length of a telephone conversation, etc.

Obviously to each value of the random variable there corresponds a definite probability and using this a *probability distribution* is defined as follows :

DEFINITION :

Probability distribution is a systematic arrangement of the possible values of a random variable and their corresponding probabilities.

For example, in tossing a balanced die once, if the variable x represents the number of spots that appears uppermost, it takes on the values 1, 2, ..., 6 with corresponding probabilities $\frac{1}{6}$ for each. Hence,

Probability distribution table associated with tossing a balanced die.

No. of spots (x)	1	2	3	4	5	6
Probability	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Also the following is the probability distribution table associated with the tossing of a fair coin thrice.

Probability distribution table associated with tossing a fair coin thrice.

No. of Heads (x)	0	1	2	3
Probability	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

In this example, the probability of getting no head is

$$^3C_0(\tfrac{1}{2})^0(\tfrac{1}{2})^3 = \tfrac{1}{8}$$

the probability of getting one head is, $^3C_1(\frac{1}{2})^1(\frac{1}{2})^{3-1} = \frac{3}{8}$

the probability of getting two heads is, $^3C_2(\frac{1}{2})^2(\frac{1}{2})^{3-2} = \frac{3}{8}$

and the probability of getting three heads is, $^3C_3(\frac{1}{2})^3(\frac{1}{2})^{3-3} = \frac{1}{8}$.

Mathematical Expectation :

Let x be a random variable with possible values $x_1, x_2, \cdots, x_n$ with corresponding probabilities $p_1, p_2, \cdots, p_n$ then the expected value of the random variable or the mathematical expectation of the random variable E(x) is defined as

$$p_1x_1+p_2x_2+\cdots+p_nx_n$$

Thus, $\text{E}(x)=p_1x_1+p_2x_2+\cdots+p_nx_n.$

The expected value E(x) is also called the *mean* of x and is denoted by $\bar{x}$ or m.

Some Properties of Expected Values :

(1) The expected value of a constant is the constant itself, *i.e.*, $\text{E}(a)=a$.

(2) $\text{E}(a+bx)=a+b\,\text{E}(x)$, where a, b are constants,

for, $\text{E}(a+bx)=p_1(a+bx_1)+p_2(a+bx_2)+\cdots+p_n(a+bx_n)$

$$=a(p_1+p_2+\cdots+p_n)+b(x_1p_1+x_2p_2+\cdots+x_np_n)$$
$$=a+b\,\text{E}(x) \text{ since, } p_1+\cdots+p_n=1.$$

(3) $\text{E}(a+bx)^2=\text{E}(a^2+2bx+b^2x^2)$

$$=p_1(a^2+2bx_1+b^2x_1^2)+p_2(a^2+2bx_2+b^2x_2^2)+\cdots$$
$$\cdots+p_n(a^2+2bx_n+b^2x_n^2)$$
$$=a^2(p_1+p_2+\cdots+p_n)+2b(p_1x_1+p_2x_2+\cdots+p_nx_n)$$
$$+b^2(p_1x_1^2+p_2x_2^2+\cdots+p_nx_n^2)$$
$$=a^2+2b\,\text{E}(x)+b^2\text{E}(x^2).$$

(4) If x and y be two random variables, then

$$\text{E}(x\pm y)=\text{E}(x)\pm\text{E}(y).$$

(5) If x and y be two *independent* random variables, then

$$\text{E}(xy)=\text{E}(x).\text{E}(y).$$

Variance of a Random Variable :

Let x be a random variable with mean $\text{E}(x)=m$. Then the variance of x is defined as

$$\text{Var }(x)=\text{E}[(x-\text{E}(x))^2]=\text{E}[(x-m)^2]$$
$$=(x_1-m)^2p_1+(x_2-m)^2p_2+\cdots+(x_n-m)^2p_n$$
$$=(x_1^2-2x_1m+m^2)p_1$$
$$+(x_2^2-2x_2m+m^2)p_2+\cdots+(x_n^2-2x_nm+m^2)p_n$$

$$=(x_1{}^2p_1+x_2{}^2p_2+\cdots+x_n{}^2p_n)$$
$$-2m(x_1p_1+x_2p_2+\cdots+x_np_n)+m^2(p_1+p_2+\cdots+p_n)$$
$$=E(x^2)-2m\,E(x)+m^2\text{ ; since }p_1+p_2+\cdots+p_n=1$$
$$=E(x^2)-2E(x).E(x)+[E(x)]^2$$
$$=E(x^2)-[E(x)]^2$$
$$=E(x^2)-m^2.$$

Thus, $\text{Var}\,(x)=E(x^2)-[E(x)]^2=E(x^2)-m^2.$

$\therefore$ $\text{S.D.}(x)=\sqrt{E(x^2)-m^2}.$

EXAMPLE :

If a coin is tossed 50 times, in how many of these tosses it is expected to find the 'head' ?

SOLUTION :

In a single toss, the probability of obtaining head is $\frac{1}{2}$.

$\therefore$ expected number of tosses with 'head' out of 50 tosses $=50\times\frac{1}{2}=25.$

EXAMPLE :

The following is the probability distribution of a random variable :

x	2	3	4	5
Probability	0·2	0·4	0·3	0·1

Find the expected value and variance of the random variable.

SOLUTION :

We have,

$$E(x)=x_1p_1+x_2p_2+x_3p_3+x_4p_4$$
$$=2\times0{\cdot}2+3\times0{\cdot}4+4\times0{\cdot}3+5\times0{\cdot}1$$
$$=0{\cdot}4+1{\cdot}2+1{\cdot}2+0{\cdot}5$$
$$=3{\cdot}3.$$

$$\begin{aligned}\text{Var}(x) &= \text{E}[\{x-\text{E}(x)\}^2]\\ &= \{x_1-\text{E}(x)\}^2p_1+\{x_2-\text{E}(x)\}^2p_2+\{x_3-\text{E}(x)\}^2p_3\\ &\qquad +\{x_4-\text{E}(x)\}^2p_4\\ &= (2-3{\cdot}3)^2\times 0{\cdot}2+(3-3{\cdot}3)^2\times 0{\cdot}4+(4-3{\cdot}3)^2\times {\cdot}03\\ &\qquad +(5-3{\cdot}3)^2\times 0{\cdot}1\\ &= 1{\cdot}69\times 0{\cdot}2+0{\cdot}9\times 0{\cdot}4+{\cdot}49\times 0{\cdot}3+2{\cdot}89\times 0{\cdot}1\\ &= 0{\cdot}338+0{\cdot}36+0{\cdot}147+0{\cdot}289\\ &= 1{\cdot}134\end{aligned}$$

[**Note.** S. D. of the random variable $=\sqrt{\text{Var}(x)}$.]

EXAMPLE :

Find the mathematical expectation of the number of points if a balanced die is rolled.

SOLUTION :

Here the random variable takes on the values 1, 2, 3, 4, 5, 6 with corresponding probabilities $\frac{1}{6}$ for each.

Hence,

$$\begin{aligned}\text{E}(x) &= 1.\tfrac{1}{6}+2.\tfrac{1}{6}+3.\tfrac{1}{6}+4.\tfrac{1}{6}+5.\tfrac{1}{6}+6.\tfrac{1}{6}\\ &= \tfrac{1}{6}(1+2+3+4+5+6)\\ &= \frac{1}{6}\cdot\frac{6\times 7}{2}=3{\cdot}5.\end{aligned}$$

EXAMPLE :

Find the mathematical expectation of receiving a tail when a balanced coin is tossed twice.

SOLUTION :

Outcomes are (H, H) ; (H, T) ; (T, H) ; (T, T). So, there is 1 way of getting two tails, 2 ways of getting one tail and 1 way of getting no tail.

Now, P (No tail) $= {\cdot}5\times{\cdot}5={\cdot}25$;

*P (One tail) $= {\cdot}5\times{\cdot}5+{\cdot}5\times{\cdot}5={\cdot}50$;

P (Two tails) $= {\cdot}5\times{\cdot}5={\cdot}25$.

If the variable x represents the number of tails it takes the values 0, 1 and 2 with corresponding probabilities ·25, ·5 and ·25. Hence,

$$\begin{aligned}\text{E}(x) &= 0\times{\cdot}25+1\times{\cdot}5+2\times{\cdot}25\\ &= 1.\end{aligned}$$

* The probability of getting one tail is the sum of the probabilities of HT, TH, and similarly the other probabilities.

EXAMPLE :

Find the expected value and the variance of number of points in rolling two balanced dice.

SOLUTION :

Here the random variable takes on the values 2, 3, 4, ..., 12 with corresponding probabilities $\frac{1}{36}, \frac{2}{36}, \frac{3}{36}, \frac{4}{36}, \frac{5}{36}, \frac{6}{36}, \frac{5}{36}, \frac{4}{36}, \frac{3}{36}, \frac{2}{36}, \frac{1}{36}$.

$$\therefore \quad E(x) = 2\times\frac{1}{36} + 3\times\frac{2}{36} + 4\times\frac{3}{36} + 5\times\frac{4}{36} + 6\times\frac{5}{36} + 7\times\frac{6}{36}$$
$$+ 8\times\frac{5}{36} + 9\times\frac{4}{36} + 10\times\frac{3}{36} + 11\times\frac{2}{36} + 12\times\frac{1}{36} = \frac{252}{36} = 7.$$

Variance :

$$\text{Var }(x)[\sigma^2] = E\,[\{x - E(x)\}^2]$$
$$= (2-7)^2\times\frac{1}{36} + (3-7)^2\times\frac{2}{36} + (4-7)^2\times\frac{3}{36} + (5-7)^2\times\frac{4}{36}$$
$$+ (6-7)^2\times\frac{5}{36} + (7-7)^2\times\frac{6}{36} + (8-7)^2\times\frac{5}{36} + (9-7)^2\times\frac{4}{36}$$
$$+ (10-7)^2\times\frac{3}{36} + (11-7)^2\times\frac{2}{36} + (12-7)^2\times\frac{1}{36}$$
$$= \frac{25}{36} + \frac{32}{36} + \frac{27}{36} + \frac{16}{36} + \frac{5}{36} + \frac{5}{36} + \frac{16}{36} + \frac{27}{36} + \frac{32}{36} + \frac{25}{36}$$
$$= \frac{210}{36} = 5{\cdot}833.$$

EXAMPLE :

A man is to play a game as follows :

In three tosses of a balanced coin, he will get a reward of Rs. 20,000, Rs. 10,000, Rs. 1,000 and no reward if he gets three tails, two tails, one tail and no tail respectively. The entrance fee for the contest is Rs. 6,000. Will he play the game ?

SOLUTION :

He will like to play the game if he receives more than Rs. 6,000 (the entrance fee).

No. of tails	Pay off (x)	Probability (p)	Expected value (px)
0	0	·5 × ·5 × ·5 = ·125	0
1	5,000	*3(·5 × ·5 × ·5) = ·375	1875
2	10,000	3(·5 × ·5 × ·5) = ·375	3750
3	20,000	·5 × ·5 × ·5 = ·125	2500
			8125

His expected return is Rs. 8,125 and since this expected return is more than Rs. 6,000, the entrance fee, he will play the game.

*The probability of getting exactly one tail is the sum of the probabilities of H H T, H T H and T H H, and similarly the other probabilities.

Note. By expected return we mean if he play long enough then on an average he may get Rs. 8,125.

Miscellaneous Examples

EXAMPLE :

A bag contains 8 red balls and 5 white balls. Two successive draws of 3 balls are made *without replacement*. Find the probability that the first drawing will give 3 white balls and the second 3 red balls. (C. A. 1978)

SOLUTION :

Total number of balls $= 8 + 5 = 13$.

3 balls can be drawn from 13 balls in ${}^{13}C_3 = 286$ ways.

So, the sample space consists of 286 sample points.

Let, $A = \{$Three balls drawn in 1st drawing are white$\}$

and $B = \{$Three balls drawn in 2nd drawing are red$\}$.

The event A consists of ${}^{5}C_3 = 10$ sample points

$\therefore \quad P(A) = \frac{10}{286}$.

Now, since the 3 white balls drawn in the 1st drawing are not replaced, the sample space now reduces to ${}^{10}C_3 = 120$ sample points, as there are now only 8 red and 2 white balls left and 3 balls can be drawn from 10 balls in ${}^{10}C_3$ ways.

The event B consists of ${}^{8}C_3 = 56$ sample points as 3 red balls can be drawn from 8 red balls in ${}^{8}C_3$ ways.

$\therefore \quad P(B/A) = \frac{56}{120}$.

$\therefore$ by theorem of compound probability,

$$P(A \cap B) = P(A).P(B/A)$$

$$= \frac{10}{286} \times \frac{56}{120} = \frac{7}{429}.$$

EXAMPLE :

A bag contains 8 red balls and 5 white balls. Two successive drawings of 3 balls are made *with replacement*. Find the probability that the first drawing will give 3 white balls and the second 3 red balls.

SOLUTION :

Total number of balls $= 8 + 5 = 13$.

Out of 13 balls, 3 balls can be drawn in ${}^{13}C_3 = 286$ ways.

So, the sample space consists of 286 sample points.

Let A = {Three balls drawn in first drawing are white}

and B = {Three balls drawn in second drawing are red}.

Now, out of 5 white balls, 3 can be drawn in $^5C_3 = 10$ ways.

∴ the event A consists of 10 sample points and hence $P(A) = \frac{10}{286}$.

Since 3 white balls drawn in the first drawing are replaced the events A and B are independent and the event B consists of $^8C_3 = 56$ sample points.

∴ $P(B) = \frac{56}{286}$.

∴ by the multiplication rule,

$$P(A \cap B) = P(A).P(B) = \frac{10}{286} \times \frac{56}{286} = \frac{140}{20449}.$$

EXAMPLE :

An article manufactured by a company consists of two parts A and B. In the process of manufacture of part A, 9 out of 100 are likely to be defective. Similarly 5 out of 100 are likely to be defective in the manufacture of part B. Calculate the probability that the assembled part will not be defective.

SOLUTION :

Let A and B denote the events that part A and part B of the article are defective respectively. Then,

$P(A) = \frac{9}{100}$ and $P(B) = \frac{5}{100}$.

The probability that the part-A will not be defective is

$P(\bar{A}) = 1 - P(A) = 1 - \frac{9}{100} = \frac{91}{100}$.

Similarly, the probability that the part-B will not be defective is

$$P(\bar{B}) = 1 - P(B) = 1 - \frac{5}{100} = \frac{95}{100}.$$

The two events A and B are independent and consequently the two complementary events $\bar{A}$ and $\bar{B}$ are independent. Hence by the multiplication rule of the probability,

$$P(\bar{A} \cap \bar{B}) = P(\bar{A}).P(\bar{B}) = \frac{81}{100}.\frac{95}{100} = \frac{8645}{10000} = 0{\cdot}8645.$$

EXAMPLE :

The probability that X can solve a problem in Business Statistics is $\frac{3}{4}$, that Y can solve it is $\frac{2}{3}$, that Z can solve it is $\frac{5}{6}$. If they all try independently, find the probability that the problem will be solved.

SOLUTION :

Let A, B and C denote the events that the problem in Business Statistics is solved by the students X, Y and Z respectively. Then

$$P(A)=\tfrac{3}{4},\ P(B)=\tfrac{2}{5} \text{ and } P(C)=\tfrac{5}{9}.$$

The problem will be solved if either A or B or C solves it. Hence, by addition rule of probability,

$$\begin{aligned}
P(A\cup B\cup C) &= P(A)+P(B)+P(C)-P(A\cap B)-P(B\cap C)-P(C\cap A)\\
&\qquad +P(A\cap B\cap C)\\
&= P(A)+P(B)+P(C)-P(A).P(B)-P(B).P(C)\\
&\qquad -P(C).P(A)+P(A).P(B).P(C)\\
&\qquad (\because \text{ events A, B, C are independent})\\
&= \tfrac{3}{4}+\tfrac{2}{5}+\tfrac{5}{9}-\tfrac{3}{4}.\tfrac{2}{5}-\tfrac{2}{5}.\tfrac{5}{9}-\tfrac{5}{9}.\tfrac{3}{4}+\tfrac{3}{4}.\tfrac{2}{5}.\tfrac{5}{9}\\
&= \tfrac{168}{180}.
\end{aligned}$$

ALTERNATIVE METHOD :

$$\begin{aligned}
\text{We have, } P(A\cup B\cup C) &= 1-P(\bar{A}\cap\bar{B}\cap\bar{C})\\
&= 1-P(\bar{A}).P(\bar{B}).P(\bar{C}) \qquad (\because \text{ A, B, C are}\\
&\qquad \text{independent events, so also } \bar{A}, \bar{B}, C)\\
&= 1-(1-P(A))\,(1-P(B))\,(1-P(C))\\
&= 1-(1-\tfrac{3}{4})\,(1-\tfrac{2}{5})\,(1-\tfrac{5}{9})\\
&= 1-\tfrac{1}{4}.\tfrac{3}{5}.\tfrac{4}{9}=\tfrac{168}{180}.
\end{aligned}$$

EXAMPLE :

A bag contains 5 red and 4 black balls. A ball is drawn at random from the bag and put into another bag which contains 3 red and 7 black balls. A ball is drawn randomly from the second bag. What is the probability that it is red ? (C. U. 1970)

SOLUTION :

Let A = event of transferring a red ball from first bag.
B = event of transferring a black ball from first bag.
C = event of drawing a red ball from second bag.

To draw a red ball from the second bag we should have either

(i) a red ball is transferred from the first bag to the second bag *and* a red ball is drawn from it.

or, (ii) a black ball is transferred from the first bag to the second bag *and* a red ball is drawn from it.

That is, we should have either (A and C) or (B and C), *i.e.*, either $(A \cap C)$ or $(B \cap C)$.

Now, since the events $A \cap C$ and $B \cap C$ are *mutually exclusive*, the required probability is

$$\begin{aligned} P[(A \cap C) \text{ or } (B \cap C)] &= P[(A \cap C) \cup (B \cap C)] \\ &= P(A \cap C) + P(B \cap C) \\ &= P(A).P(C \mid A) + P(B).P(B \mid C). \end{aligned}$$

Now, $P(A) = \dfrac{5}{5+4} = \dfrac{5}{9}$, $P(B) = \dfrac{4}{5+4} = \dfrac{4}{9}$

$P(C \mid A) = \frac{4}{11}$ (since a red ball is transferred from 1st bag to the 2nd bag).

$P(C \mid B) = \frac{3}{11}$ (Since a black ball is tranferred from 1st bag to the 2nd bag).

$\therefore$ required Probability $= \frac{5}{9}.\frac{4}{11} + \frac{4}{9}.\frac{3}{11} = \frac{32}{99}$.

EXAMPLE :

An urn contains 5 red and 4 black balls and another urn contains 3 red and 7 black balls. If one ball is drawn from each urn, find the probability that (i) both are of same colour and (ii) both are of different colours.

SOLUTION :

Let R_1 = {Ball drawn from 1st urn is red},

B_1 = {Ball drawn from 1st urn is black},

R_2 = {Ball drawn from 2nd urn is red},

B_2 = {Ball drawn from 2nd urn is black}.

1st part : When both the balls are of same colour.

To obtain the balls of the same colour we should have either both balls red or both balls black, *i.e.*, either $(R_1 \cap R_2)$ or $(B_1 \cap B_2)$. Now, since the events $R_1 \cap R_2$ and $B_1 \cap B_2$ are mutually exclusive, we have by addition rule of probability,

$$\begin{aligned} P[(R_1 \cap R_2) \text{ or } (B_1 \cap B_2)] &= P[(R_1 \cap R_2) \cup (B_1 \cap B_2)] \\ &= P[(R_1 \cap R_2) + P(B_1 \cap B_2) \qquad \cdots \quad (1) \end{aligned}$$

Again the events R_1 and R_2 are independent and so also the events B_1 and B_2, since drawing a ball from 1st urn does not affect the drawing of a ball from the 2nd urn.

$\therefore \quad P(R_1 \cap R_2) = P(R_1).P(R_2),\ P(B_1 \cap B_2) = P(B_1).P(B_2)$

Hence from (1) the required probability $= P(R_1).P(R_2) + P(B_1).P(B_2)$

$$= \tfrac{5}{9}.\tfrac{3}{10} + \tfrac{4}{9}.\tfrac{7}{10}$$

$$= \tfrac{43}{90}$$

2nd part : Proceeding in the same way as above, the required probability in this case is :

$$P[(R_1 \cap B_2) \text{ or } (B_1 \cap R_2)] = P[(R_1 \cap B_2) \cup (B_1 \cap R_2)]$$

$$= P(R_1 \cap B_2) + (B_1 \cap R_2)$$

$$= P(R_1).P(B_2) + P(B_1).P(R_2)$$

$$= \tfrac{5}{9}.\tfrac{7}{10} + \tfrac{4}{9}.\tfrac{3}{10}$$

$$= \tfrac{47}{90}.$$

EXAMPLE

Two persons X and Y appear in an interview for two vacancies in the same post. The probability of X's selection is $\frac{2}{11}$ and that of Y's selection is $\frac{1}{7}$. What is the probability that (i) both of them will be selected, (ii) only one of them will be selected, and (iii) none of them will be selected.

SOLUTION :

Let A denote the event that X is selected and B denote the event that Y is selected. Then clearly the events A and B are independent.

(i) Probability that both of them will be selected

$$= P(A \text{ and } B)$$

$$= P(A \cap B)$$

$$= P(A).P(B)$$

$$= \tfrac{2}{11}.\tfrac{1}{7} = \tfrac{2}{77}.$$

(ii) Probability that one will be selected

= P[(X will be selected *and* Y will not be selected) or (X will not be selected *and* Y will be selected)]

$= P[(A \cap \bar{B}) \cup (\bar{A} \cap B)]$

$= X(A \cap \bar{B}) + P(\bar{A} \cap B)$, (since $A \cap \bar{B}$ and $\bar{A} \cap B$ are mutually exclusive)

$= P(A).P(\bar{B}) + P(\bar{A}).P(B).$

$= P(A).[1 - P(B)] + [1 - P(A)].P(B)$

$= \tfrac{2}{11}.\tfrac{6}{7} + \tfrac{9}{11}.\tfrac{1}{7}$

$= \tfrac{21}{77}.$

(iii) Probability that none will be selected

$= P(\bar{A} \cap \bar{B}) = P(\bar{A}).P(\bar{B}) = [1 - P(A)].[1 - P(B)]$

$= (1 - \frac{2}{11}).(1 - \frac{1}{7}) = \frac{9}{11}.\frac{6}{7} = \frac{54}{77}.$

EXERCISE 14

1. An urn contains 8 white and 3 red balls. If two balls are drawn at random, find the probability that (i) both are white, (ii) both are red, (iii) one is of each colour. (C. U. 1973) [$\frac{28}{55}, \frac{3}{55}, \frac{24}{55}$]

2. What is the chance of picking a spade or an ace not of spade from a pack of 52 cards ? (C. U. 1965) [$\frac{4}{13}$]

3. Four balls are drawn at random from a bag containing 5 white and 7 black balls. Find the probability of getting (i) 4 white balls, (ii) 2 white and 2 black balls, (iii) 3 black and 1 white ball. [$\frac{1}{99}, \frac{14}{33}, \frac{35}{99}$]

4. There are 3 Geologists, 4 Engineers, 2 Statisticians and 1 Doctor. A committee of 4 from among them is to be formed. Find the probability that the committee—(i) consists of one of each kind, (ii) has at least one Geologist, (iii) has the Doctor as a member and three others. [$\frac{4}{35}, \frac{5}{6}, \frac{2}{5}$]

5. Two dice are thrown. Find the probability that (i) the first die shows 4, (ii) the total of the numbers on the dice is 9 or greater than 9, (iii) the number on the first die is greater than the number of the second die. [$\frac{1}{6}, \frac{5}{18}, \frac{5}{12}$]

6. An urn contains 3 red, 4 white and 5 black balls. Three balls are drawn at random from the urn. Find the probability that (i) all are black, (ii) all are of different colours. [$\frac{1}{22}, \frac{3}{11}$]

7. The odds against X solving a problem are 8 to 6 and the odds in favour of Y solving the same problem are 14 to 10. What is the probability that if they both try, the problem will be solved at least by one of them ? [$\frac{16}{21}$]

8. An urn A contains 2 white and 4 black balls. Another urn B contains 5 white and 7 black balls. A ball is transferred from urn A to the urn B. Then a ball is drawn from urn B. Find the probability that it will be white. [$\frac{16}{39}$]

9. A, B and C, in that order, toss a coin. The first one to throw a head wins. What are their respective chances of winning. Assume that the game may continue indefinitely. [$\frac{4}{7}, \frac{2}{7}, \frac{1}{7}$]

10. The probability that a candidate passes in Business Mathematics is 0·60 and that he passes in Business Statistics is 0·50. What is the probability that the candidate passes only in any one of the two subjects. [$\frac{1}{2}$]

11. The incidence of occupational disease is such that on the average 20% of workers suffer from it. If 10 workers are selected at random, find the probability that (i) exactly 2 workers suffer from the disease, (ii) not more than two workers suffer from the disease.

(C. U. 1974) [0·302, 0·678]

12. Two dice are thrown n times in succession. What is the probability of obtaining double-six at least once ?

(C. U. 1975) [$1-(\frac{35}{36})^n$]

13. A problem in Statistics is given to three students A, B, C whose chances of solving it are $\frac{1}{3}$, $\frac{1}{4}$, $\frac{1}{5}$ respectively. What is the probability that the problem will be solved ? [$\frac{3}{5}$]

14. There are two identical boxes containing respectively 4 white and 3 red balls, and 3 white and 7 red balls. A box is chosen at random and a ball is drawn from it. Find the probability that the ball is white. (C. U. 1976) [$\frac{61}{140}$]

15. A person is known to hit the target 3 out of 4 shots, whereas another person is known to hit 2 out of 3 shots. Find the probability that the target being hit when they both try. [$\frac{11}{12}$]

16. The probability that a student passes in a Physics test is $\frac{2}{3}$ and the probability that he passes both a Physics test and an English test is $\frac{14}{45}$. The probability that he passes at least one test is $\frac{4}{5}$. What is the probability that he passes the English test ? [$\frac{4}{9}$]

17. A card is drawn at random from a pack of 52 cards. If ace counts one, king, queen and jack count 10 each and others count at their face value, show that the expectation of the value of the card is $\frac{85}{13}$.

18. In a group of equal number of men and women, 10% men and 45% women are unemployed. What is the probability that a person selected at random is unemployed ? [$\frac{11}{40}$]

19. A purse contains 3 silver coins and 4 copper coins and a second purse contains 4 silver and 3 copper coins. If a coin is selected at random from one of the two purses, what is the probability that it is a silver coin ? (C. U. 1968) [$\frac{1}{2}$]

20 The probability that an entering college student will be a graduate is 0·4. Determine the probability that out of 5 entering students, (i) none, (ii) one, (iii) at least one, will be a graduate.

(C. U. 1964) [·07776, ·2592, ·92224]

21. In a given business venture a man can make a profit of Rs. 300 with probability 0·6 or incur a loss of Rs. 100 with probability 0·4. Calculate his expectation. (C. U. 1966) [Rs. 140]

22. A can solve 75% of the problems in this book and B can solve 70%. What is the probability that either A or B can solve a problem chosen at random ? [$\frac{37}{40}$]

23. A packet of 10 electronic components is known to include 3 defectives. If 4 components are randomly chosen and tested, what is the probability of finding among them not more than one defective ?

(C. U. 1980) [0·6517]

24. Two dice are thrown, find the expected value for the sum of their face numbers. [7]

25. Find the expected value of the product of points on two dice. [$\frac{49}{4}$]

UNIVERSITY QUESTIONS

CALCUTTA UNIVERSITY

BUSINESS STATISTICS [Honours]

1980

Group A

1. (*a*) Draw up a blank table to show the number of employees in a large commercial firm, classified according to (*i*) Sex : male and female ; (*ii*) Three age-groups : below 30, 30 and above but below 45, 45 and above ; and (*iii*) Four income-groups : below Rs. 400, Rs. 400 — 750, Rs. 750 — 1,000, above Rs. 1,000.

(*b*) The following data show the estimated savings of the household sector in India during 1962-63, as revealed by the C.S.O.

Form of Savings	*Amount (Rs. crores)*
Currency ...	... 175
Provident Fund ...	... 145
Physical ...	... 158
Others ...	... 440

Present the information in a suitable diagram so as to enable comparison among the various components and also in relation to the total.

2. (*a*) What are 'quartiles' of a distribution ? How do you use them for measuring dispersion ?

(*b*) Calculate the mean and the median from the following data :

Weekly Wages (Rs.)		*Number of Workers*
Below 10	...	8
,, 20	...	18
,, 30	...	45
,, 40	...	90
,, 50	...	113
,, 60	...	120

3. (*a*) Let $x_1, x_2, \ldots, x_n$ be the values of a variable with frequencies $f_1, f_2, \ldots, f_n$ respectively. Prove that

$$\sum_{i=1}^{n} f_i\,(x_i-\bar{x})=0.$$

(*b*) For a group of 50 boys the mean score and the standard deviation of scores on a test are 59·5 and 8·38 respectively. For a group of 40 girls the same results are 54·0 and 8·23 respectively. Find the mean and the standard deviation of the combined group of 90 children

4. (*a*) Define correlation coefficient and state its important properties (Clearly explain all the symbols you use).

(*b*) A sample of size $n=16$ yield the following sums :

$\Sigma x=749$, $\Sigma y=77{\cdot}90$, $\Sigma y^2=454{\cdot}81$, $\Sigma xy=3156{\cdot}80$ and $\Sigma x^2=42{\cdot}177$.

Compute the linear regression equation of x on y.

5. (*a*) Discuss the merits and limitations of the moving average method for determining the trend in the analysis of time series.

(*b*) The following series of observations is known to have a business cycle with a period of 4 years. Find the trend values by the moving average method.

Year	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980
Production ('000 tons)	506	620	1036	673	588	696	1116	738	663	773	1189

6. (*a*) Discuss the different steps that have to be taken in the construction of a price index number.

(*b*) The values of a function $f(x)$ are given below for some specified values of x :

x	3	4	5	9
$f(x)$	6	5	-2	30

Using an appropriate interpolation formula, find the value of $f(7)$.

Group B

7. (*a*) Define the terms : Null set, Disjoint sets, Finite and infinite sets, Complement of a set. Give *one* example for each.

(*b*) Prove that

$$A \cup (B \cap C)=(A \cup B) \cap (A \cup C).$$

8. (*a*) Define and illustrate the following terms :

Mutually exclusive events, Exhaustive set of events, Independent events.

(*b*) Boxes I and II contain respectively 4 white, 3 red and 3 blue balls ; and 5 white, 4 red and 3 blue balls. If one ball is drawn at random from each box, what is the probability that both the balls are of the same colour ?

9. (*a*) A and B are two events, not mutually exclusive, connected with a random experiment E. If $P(A)=1/4$, $P(B)=2/5$ and $P(A \cup B)=1/2$, find the values of the following probabilities :—

(*i*) $P(A \cap B)$, (*ii*) $P(A \cap B^c)$, (*iii*) $P(A^c \cup B^c)$

where c stands for the complement.

(*b*) A bag contains 7 red balls and 5 white balls. 4 balls are drawn at random. What is the probability that (*i*) all of them are red ; (*ii*) two of them would be red and two white ?

10. (*a*) How do you distinguish between 'discrete' and 'continuous' random variables ? Illustrate your answer with suitable examples.

(*b*) A random variable has the following probability distribution :

x	4	5	6	8
Probability	0·1	0·3	0·4	0·2

Find the expectation and the standard deviation of the random variable.

11. (*a*) State and prove the *Addition Theorem* of probability for two mutually exclusive events.

(*b*) A packet of 10 electronic components is known to include 3 defectives. If 4 components are randomly chosen and tested, what is the probability of finding among them more than one defective ?

12. Write notes on (*any three*) :

(*a*) Universal set and subset,

(*b*) Classical concept of probability,

(*c*) Random variable,

(*d*) Dependent and Independent events.

1981

1. (*a*) Discuss briefly different types of diagrams generally used to represent numerical data and point out the relative merits and demerits of diagrammatic representation compared to other methods used for the purpose.

(*b*) Draw the bistogram of the following frequency distribution and use it to find the total number of wage-earners in the age group 19-32 years :

Age-group	14-15	16-17	18-20	21-24	25-29	30-34	35-39
No. of wage-earners	60	140	150	110	110	100	90

2. (*a*) Indicate merits and shortcomings of the different cetral tendency as well as situations where to use each.

(*b*) In the following frequency distribution, two class-frequencies are missing :

Intelligence Quotient	No. of students	Intelligence Quotient	No. of students
55— 64	2	105—114	?
65— 74	19	115—124	92
75— 84	78	125—134	14
85— 94	?	135—144	4
95—104	301		

It is however known that the total frequency is 900 and the Median 100·048. Find the two missing frequencies.

3. (*a*) Prove that the sum of squares of deviations is minimum when deviations are measured from the mean.

(*b*) From the following table calculate the values of (*i*) Mean, (*ii*) Standard deviation and (*iii*) Coefficient of variation :

Monthly wages	No. of servants	Monthly wages	No. of servants
0—10	1	50—60	35
10—20	4	60—70	10
20—30	10	70—80	7
30—40	22	80—90	1
40—50	30		

4. (*a*) Show that product-moment correlation coefficient 'r' lies between +1 and −1.

(*b*) The following data give the height in inches (x) and the weight in *lb* (y) of 10 students of age 17 years :

x	61	66	68	64	65	70	63	62	64	67
y	112	123	130	115	110	125	100	113	116	126

Calculate the correlation coefficient.

5. (*a*) Explain clearly what is meant by Time series analysis. Describe briefly the various characteristic movements of a time series.

(*b*) Construct Indices of seasonal variation from the following time series data on consumption of cold drinks (in 1,000 bottles) :

Quarter / Year	I	II	III	IV
1971	90	75	87	70
1972	75	80	78	75
1973	80	75	75	72
1974	85	82	80	81

6. (*a*) Prepare price and quantity index numbers for 1972 with 1961 as base year from the following data by using (*i*) Laspeyre's, (*ii*) Paache's and (*iii*) Fisher's method.

Commodity	Unit	1961		1972	
		Quantity	Price (Rs.)	Quantity	Price (Rs.)
A	Kg	5	2·00	7	4·50
B	Quintal	7	2·50	10	3·20
C	Dozen	6	8·00	6	4·50
D	Kg	2	1·00	9	1·80

(*b*) The following table gives values of an unknown function $f(x)$ for certain equidistant values of x. Use a suitable interpolation formula to find the value of the function at $x = 24$:

x	18	22	26	30	34
$f(x)$	11·725	14·221	16·958	19·954	23·222

7. (*a*) Let A and B be any two sets, prove that

$$(A \cup B)' = A' \cap B'$$

Illustrate the theorem by Venn diagram.

(*b*) Let $A = \{a, b, c\}$, $B = \{a, b\}$, $C = \{a, b, d\}$, $D = \{c, d\}$ and $E = \{d\}$

State which of the following statements are correct and give reasons

(*i*) $B \subset A$ (*ii*) $D \not\supset E$ (*iii*) $D \subset B$ (*iv*) $\{a\} \subset A$

8. (*a*) Let A and B be two independent events. Then show that (*i*) A and B^c, (*ii*) A^c and B^c are also independent.

(*b*) There are three men aged 60, 65 and 70 years. The probability to live 5 years more is 0·8 for a 60 year old, 0·6 for a 65 year old and 0·3 for a 70 year old person. Find the probability that at least two of the three persons will remain alive 5 years hence.

9. (*a*) What is meant by compound event in probability ? State and prove the theorem of compound probability.

(*b*) A number is chosen at random from the set 1, 2, 3, ... , 100 and another number is chosen at random from the set 1, 2, ... , 50. What is the expected value of the product ?

10. Write notes on (*any three*) :

(*a*) Frequency interpretation of probability.

(*b*) Union, intersection and difference of two events.

(*c*) Random variable.

(*d*) Ordered pair and Cartesian product.

LOGARITHMIC TABLES

LOGARITHMS

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
10	0000	0043	0086	0128	0170						5	9	13	17	21	26	30	34	38
						0212	0253	0294	0334	0374	4	8	12	16	20	24	28	32	36
11	0414	0453	0492	0531	0569						4	8	12	16	20	23	27	31	35
						0607	0645	0682	0719	0755	4	7	11	15	18	22	26	29	33
12	0792	0828	0864	0899	0934						3	7	11	14	18	21	25	28	32
						0969	1004	1038	1072	1106	3	7	10	14	17	20	24	27	31
13	1139	1173	1206	1239	1271						3	6	10	13	16	19	23	26	29
						1303	1335	1367	1399	1430	3	7	10	13	16	19	22	25	29
14	1461	1492	1523	1553	1584						3	6	9	12	15	19	22	25	28
						1614	1644	1673	1703	1732	3	6	9	12	14	17	20	23	26
15	1761	1790	1818	1847	1875						3	6	9	11	14	17	20	23	26
						1903	1931	1959	1987	2014	3	6	8	11	14	17	19	22	25
16	2041	2068	2095	2122	2148						3	6	8	11	14	16	19	22	24
						2175	2201	2227	2253	2279	3	5	8	10	13	16	18	21	23
17	2304	2330	2355	2380	2405						3	5	8	10	13	15	18	20	23
						2430	2455	2480	2504	2529	3	5	8	10	12	15	17	20	22
18	2553	2577	2601	2625	2648						2	5	7	9	12	14	17	19	21
						2672	2695	2718	2742	2765	2	4	7	9	11	14	16	18	21
19	2788	2810	2833	2856	2878						2	4	7	9	11	13	16	18	20
						2900	2923	2945	2967	2989	2	4	6	8	11	13	15	17	19
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	2	4	6	8	11	13	15	17	19
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	2	4	6	8	10	12	14	16	18
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598	2	4	6	8	10	12	14	15	17
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784	2	4	6	7	9	11	13	15	17
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962	2	4	5	7	9	11	12	14	16
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133	2	3	5	7	9	10	12	14	15
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298	2	3	5	7	8	10	11	13	15
27	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456	2	3	5	6	8	9	11	13	14
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609	2	3	5	6	8	9	11	12	14
29	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757	1	3	4	6	7	9	10	12	13
30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900	1	3	4	6	7	9	10	11	13
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038	1	3	4	6	7	8	10	11	12
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172	1	3	4	5	7	8	9	11	12
33	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302	1	3	4	5	6	8	9	10	12
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428	1	3	4	5	6	8	9	10	11
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551	1	2	4	5	6	7	9	10	11
36	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670	1	2	4	5	6	7	8	10	11
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786	1	2	3	5	6	7	8	9	10
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899	1	2	3	5	6	7	8	9	10
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010	1	2	3	4	5	7	8	9	10
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117	1	2	3	4	5	6	8	9	10
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222	1	2	3	4	5	0	7	8	9
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325	1	2	3	4	5	6	7	8	9
43	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425	1	2	3	4	5	6	7	8	9
44	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522	1	2	3	4	5	6	7	8	9
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618	1	2	3	4	5	6	7	8	9
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712	1	2	3	4	5	6	7	7	8
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803	1	2	3	4	5	5	6	7	8
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893	1	2	3	4	4	5	6	7	8
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981	1	2	3	4	4	5	6	7	8

LOGARITHMS

	0	1	2	3	4	5	6	7	8	9	123	456	789
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067	1 2 3	3 4 5	6 7 8
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152	1 2 3	3 4 5	6 7 8
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235	1 2 2	3 4 5	6 7 7
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316	1 2 2	3 4 5	6 6 7
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396	1 2 2	3 4 5	6 6 7
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474	1 2 2	3 4 5	5 6 7
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551	1 2 2	3 4 5	5 6 7
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627	1 2 2	3 4 5	5 6 7
58	7034	7642	7649	7657	664	7672	7679	7686	7694	7701	1 1 2	3 4 4	5 6 7
59	7709	7716	7723	7731	738	7745	7752	7760	7767	7774	1 1 2	3 4 4	5 6 7
60	7782	7789	7796	7803	810	7818	7825	7832	7839	7846	1 1 2	3 4 4	5 6 6
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917	1 1 2	3 4 4	5 6 6
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987	1 1 2	3 3 4	5 6 6
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055	1 1 2	3 3 4	5 5 6
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122	1 1 2	3 3 4	5 5 6
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189	1 1 2	3 3 4	5 5 6
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254	1 1 2	3 3 4	5 5 6
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319	1 1 2	3 3 4	5 5 6
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382	1 1 2	3 3 4	4 5 6
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445	1 1 2	2 3 4	4 5 6
70	8451	8457	8463	8470	8476	8482	8488	8494	8500	8506	1 1 2	2 3 4	4 5 6
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567	1 1 2	2 3 4	4 5 5
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627	1 1 2	2 3 4	4 5 5
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686	1 1 2	2 3 4	4 5 5
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745	1 1 2	2 3 4	4 5 5
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802	1 1 2	2 3 3	4 5 5
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859	1 1 2	2 3 3	4 5 5
77	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915	1 1 2	2 3 3	4 4 5
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971	1 1 2	2 3 3	4 4 5
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025	1 1 2	2 3 3	4 4 5
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079	1 1 2	2 3 3	4 4 5
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133	1 1 2	2 3 3	4 4 5
82	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186	1 1 2	2 3 3	4 4 5
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238	1 1 2	2 3 3	4 4 5
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289	1 1 2	2 3 3	4 4 5
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340	1 1 2	2 3 3	4 4 5
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390	1 1 2	2 3 3	4 4 5
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440	0 1 1	2 2 3	3 4 4
88	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489	0 1 1	2 2 3	3 4 4
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538	0 1 1	2 2 3	3 4 4
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586	0 1 1	2 2 3	3 4 4
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633	0 1 1	2 2 3	3 4 4
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680	0 1 1	2 2 3	3 4 4
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727	0 1 1	2 2 3	3 4 4
94	9731	9730	9741	9745	9750	9754	9759	9763	9768	9773	0 1 1	2 2 3	3 4 4
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818	0 1 1	2 2 3	3 4 4
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863	0 1 1	2 2 3	3 4 4
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908	0 1 1	2 2 3	3 4 4
98	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952	0 1 1	2 2 3	3 4 4
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996	0 1 1	2 2 3	3 3 4

ANTILOGARITHMS

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
·00	1000	1002	1005	1007	1009	1012	1014	1016	1019	1021	0	0	1	1	1	1	2	2	2
·01	1023	1026	1028	1030	1033	1035	1038	1040	1042	1045	0	0	1	1	1	1	2	2	2
·02	1047	1050	1052	1054	1057	1059	1062	1064	1067	1069	0	0	1	1	1	1	2	2	2
·03	1072	1074	1076	1079	1081	1084	1086	1089	1091	1094	0	0	1	1	1	1	2	2	2
·04	1096	1099	1102	1104	1107	1109	1112	1114	1117	1119	0	1	1	1	1	2	2	2	2
·05	1122	1125	1127	1130	1132	1135	1138	1140	1143	1146	0	1	1	1	1	2	2	2	2
·06	1148	1151	1153	1156	1159	1161	1164	1167	1169	1172	0	1	1	1	1	2	2	2	2
·07	1175	1178	1180	1183	1186	1189	1191	1194	1197	1199	0	1	1	1	1	2	2	2	2
·08	1202	1205	1208	1211	1213	1216	1219	1222	1225	1227	0	1	1	1	1	2	2	2	3
·09	1230	1233	1236	1239	1242	1245	1247	1250	1253	1256	0	1	1	1	1	2	2	2	3
·10	1259	1262	1265	1268	1271	1274	1276	1279	1282	1285	0	1	1	1	1	2	2	2	3
·11	1288	1291	1294	1297	1300	1303	1306	1309	1312	1315	0	1	1	1	2	2	2	2	3
·12	1318	1321	1324	1327	1330	1334	1337	1340	1343	1346	0	1	1	1	2	2	2	2	3
·13	1349	1352	1355	1358	1361	1365	1368	1371	1374	1377	0	1	1	1	2	2	2	3	3
·14	1380	1384	1387	1390	1393	1396	1400	1403	1406	1409	0	1	1	1	2	2	2	3	3
·15	1413	1416	1419	1422	1426	1429	1432	1435	1439	1442	0	1	1	1	2	2	2	3	3
·16	1445	1449	1452	1455	1459	1462	1466	1469	1472	1476	0	1	1	1	2	2	2	3	3
·17	1479	1483	1486	1489	1493	1496	1500	1503	1507	1510	0	1	1	1	2	2	2	3	3
·18	1514	1517	1521	1524	1528	1531	1535	1538	1542	1545	0	1	1	1	2	2	2	3	3
·19	1549	1552	1556	1560	1563	1567	1570	1574	1578	1581	0	1	1	1	2	2	3	3	3
·20	1585	1589	1592	1596	1600	1603	1607	1611	1614	1618	0	1	1	1	2	2	3	3	3
·21	1622	1626	1629	1633	1637	1641	1644	1648	1652	1656	0	1	1	2	2	2	3	3	3
·22	1660	1663	1667	1671	1675	1679	1683	1687	1690	1694	0	1	1	2	2	2	3	3	3
·23	1698	1702	1706	1710	1714	1718	1722	1726	1730	1734	0	1	1	2	2	2	3	3	4
·24	1738	1742	1746	1750	1754	1758	1762	1766	1770	1774	0	1	1	2	2	2	3	3	4
·25	1778	1782	1786	1791	1795	1799	1803	1807	1811	1816	0	1	1	2	2	2	3	3	4
·26	1820	1824	1828	1832	1837	1841	1845	1849	1854	1858	0	1	1	2	2	3	3	3	4
·27	1862	1866	1871	1875	1879	1884	1888	1892	1897	1901	0	1	1	2	2	3	3	3	4
·28	1905	1910	1914	1919	1923	1928	1932	1936	1941	1945	0	1	1	2	2	3	3	4	4
·29	1950	1954	1959	1963	1968	1972	1977	1982	1986	1991	0	1	1	2	2	3	3	4	4
·30	1995	2000	2004	2009	2014	2018	2023	2028	2032	2037	0	1	1	2	2	3	3	4	4
·31	2042	2046	2051	2056	2061	2065	2070	2075	2080	2084	0	1	1	2	2	3	3	4	4
·32	2089	2094	2099	2104	2109	2113	2118	2123	2128	2133	0	1	1	2	2	3	3	4	4
·33	2138	2143	2148	2153	2158	2163	2168	2173	2178	2183	0	1	1	2	2	3	3	4	4
·34	2188	2193	2198	2203	2208	2213	2218	2223	2228	2234	1	1	2	2	3	3	4	4	5
·35	2239	2244	2249	2254	2259	2265	2270	2275	2280	2286	1	1	2	2	3	3	4	4	5
·36	2291	2296	2301	2307	2312	2317	2323	2328	2333	2339	1	1	2	2	3	3	4	4	5
·37	2344	2350	2355	2360	2366	2371	2377	2382	2388	2393	1	1	2	2	3	3	4	4	5
·38	2399	2404	2410	2415	2421	2427	2432	2438	2443	2449	1	1	2	2	3	3	4	4	5
·39	2455	2460	2466	2472	2477	2483	2489	2495	2500	2506	1	1	2	2	3	3	4	5	5
·40	2512	2518	2523	2529	2535	2541	2547	2553	2559	2564	1	1	2	2	3	4	4	5	5
·41	2570	2576	2582	2588	2594	2600	2606	2612	2618	2624	1	1	2	2	3	4	4	5	5
·42	2630	2636	2642	2649	2655	2661	2667	2673	2679	2685	1	1	2	2	3	4	4	5	6
·43	2692	2698	2704	2710	2716	2723	2729	2735	2742	2748	1	1	2	3	3	4	4	5	6
·44	2754	2761	2767	2773	2780	2786	2793	2799	2805	2812	1	1	2	3	3	4	4	5	6
·45	2818	2825	2831	2838	2844	2851	2858	2864	2871	2877	1	1	2	3	3	4	5	5	6
·46	2884	2891	2897	2904	2911	2917	2924	2931	2938	2944	1	1	2	3	3	4	5	5	6
·47	2951	2958	2965	2972	2979	2985	2992	2999	3006	3013	1	1	2	3	3	4	5	5	6
·48	3020	3027	3034	3041	3048	3055	3062	3069	3076	3083	1	1	2	3	4	4	5	6	6
·49	3090	3097	3105	3112	3119	3126	3133	3141	3148	3155	1	1	2	3	4	4	5	6	6

ANTILOGARITHMS

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
·50	3162	3170	3177	3184	3192	3199	3206	3214	3221	3228	1	1	2	3	4	4	5	6	7
·51	3236	3243	3251	3258	3266	3273	3281	3289	3296	3304	1	2	2	3	4	5	5	6	7
·52	3311	3319	3327	3334	3342	3350	3357	3365	3373	3381	1	2	2	3	4	5	5	6	7
·53	3388	3396	3404	3412	3420	3428	3436	3443	3451	3459	1	2	2	3	4	5	6	6	7
·54	3467	3475	3483	3491	3499	3508	3516	3524	3532	3540	1	2	2	3	4	5	6	6	7
·55	3548	3556	3565	3573	3581	3589	3597	3606	3614	3622	1	2	2	3	4	5	6	7	7
·56	3631	3639	3648	3656	3664	3673	3681	3690	3698	3707	1	2	3	3	4	5	6	7	8
·57	3715	3724	3733	3741	3750	3758	3767	3776	3784	3793	1	2	3	3	4	5	6	7	8
·58	3802	3811	3819	3828	3837	3846	3855	3864	3873	3882	1	2	3	4	4	5	6	7	8
·59	3890	3899	3908	3917	3926	3936	3945	3954	3963	3972	1	2	3	4	5	5	6	7	8
·60	3981	3990	3999	4009	4018	4027	4036	4046	4055	4064	1	2	3	4	5	6	6	7	8
·61	4074	4083	4093	4102	4111	4121	4130	4140	4150	4159	1	2	3	4	5	6	7	8	9
·62	4169	4178	4188	4198	4207	4217	4227	4236	4246	4256	1	2	3	4	5	6	7	8	9
·63	4266	4276	4285	4295	4305	4315	4325	4335	4345	4355	1	2	3	4	5	6	7	8	9
·64	4365	4375	4385	4395	4406	4416	4426	4436	4446	4457	1	2	3	4	5	6	7	8	9
·65	4467	4477	4487	4498	4508	4519	4529	4539	4550	4560	1	2	3	4	5	6	7	8	9
·66	4571	4581	4592	4603	4613	4624	4634	4645	4656	4667	1	2	3	4	5	6	7	9	10
·67	4677	4688	4699	4710	4721	4732	4742	4753	4764	4775	1	2	3	4	5	7	8	9	10
·68	4786	4797	4808	4819	4831	4842	4853	4864	4875	4887	1	2	3	4	6	7	8	9	10
·69	4898	4909	4920	4932	4943	4955	4966	4977	4989	5000	1	2	3	5	6	7	8	9	10
·70	5012	5023	5035	5047	5058	5070	5082	5093	5105	5117	1	2	4	5	6	7	8	9	11
·71	5129	5140	5152	5164	5176	5188	5200	5212	5224	5236	1	2	4	5	6	7	8	10	11
·72	5248	5260	5272	5284	5297	5309	5321	5333	5346	5358	1	2	4	5	6	7	9	10	11
·73	5370	5383	5395	5408	5420	5433	5445	5458	5470	5483	1	3	4	5	6	8	9	10	11
·74	5495	5508	5521	5534	5546	5559	5572	5585	5598	5610	1	3	4	5	6	8	9	10	12
·75	5623	5636	5649	5662	5675	5689	5702	5715	5728	5741	1	3	4	5	7	8	9	10	12
·76	5754	5768	5781	5794	5808	5821	5834	5848	5861	5875	1	3	4	5	7	8	9	11	12
·77	5888	5902	5916	5929	5943	5957	5970	5984	5998	6012	1	3	4	5	7	8	10	11	12
·78	6026	6039	6053	6067	6081	6095	6109	6124	6138	6152	1	3	4	6	7	8	10	11	13
·79	6166	6180	6194	6209	6223	6237	6252	6266	6281	6295	1	3	4	6	7	9	10	11	13
·80	6310	6324	6339	6353	6368	6383	6397	6412	6427	6442	1	3	4	6	7	9	10	12	13
·81	6457	6471	6486	6501	6516	6531	6546	6561	6577	6592	2	3	5	6	8	9	11	12	14
·82	6607	6622	6637	6653	6668	6683	6699	6714	6730	6745	2	3	5	6	8	9	11	12	14
·83	6761	6776	6792	6808	6823	6839	6855	6871	6887	6902	2	3	5	6	8	9	11	12	14
·84	6918	6934	6950	6966	6982	6998	7015	7031	7047	7063	2	3	5	6	8	10	11	13	15
·85	7079	7096	7112	7129	7145	7161	7178	7194	7211	7228	2	3	5	7	8	10	12	13	15
·86	7244	7261	7278	7295	7311	7328	7345	7362	7379	7396	2	3	5	7	8	10	12	13	15
·87	7413	7430	7447	7464	7482	7499	7516	7534	7551	7568	2	3	5	7	9	10	12	14	16
·88	7586	7603	7621	7638	7656	7674	7691	7709	7727	7745	2	4	5	7	9	11	12	14	16
·89	7762	7780	7798	7816	7834	7852	7870	7889	7907	7925	2	4	5	7	9	11	13	14	16
·90	7943	7962	7980	7998	8017	8035	8054	8072	8091	8110	2	4	6	7	9	11	13	15	17
·91	8128	8147	8166	8185	8204	8222	8241	8260	8279	8299	2	4	6	8	9	11	13	15	17
·92	8318	8337	8356	8375	8395	8414	8433	8453	8472	8492	2	4	6	8	10	12	14	15	17
·93	8511	8531	8551	8570	8590	8610	8630	8650	8670	8690	2	4	6	8	10	12	14	16	18
·94	8710	8730	8750	8770	8790	8810	8831	8851	8872	8892	2	4	6	8	10	12	14	16	18
·95	8913	8933	8954	8974	8995	9016	9036	9057	9078	9099	2	4	6	8	10	12	15	17	19
·96	9120	9141	9162	9183	9204	9226	9247	9268	9290	9311	2	4	6	8	11	13	15	17	19
·97	9333	9354	9376	9397	9419	9441	9462	9484	9506	9528	2	4	7	9	11	13	15	17	20
·98	9550	9572	9594	9616	9638	9661	9683	9705	9727	9750	2	4	7	9	11	13	16	18	20
·99	9772	9795	9817	9840	9863	9886	9908	9931	9954	9977	2	5	7	9	11	14	16	18	20